CONSIDERATIONS IN USING INDIVIDUAL SOCIOECONOMIC CHARACTERISTICS IN THE ANALYSIS OF MORTALITY

Mary Grace Kovar and James A. Weed
National Center for Health Statistics

In the last paragraph of their APHA monograph Differential Mortality in the United States, Kitagawa and Hauser (1973) gave strong support to the view that the improvement of social-economic conditions would be the most promising route to take in achieving further mortality reduction:

> Perhaps the most important next gain in mortality reduction is to be achieved through improved social-economic conditions rather than through increments to and application of biomedical knowledge. Certainly the biomedical know-how now available is either not available to the lower socioeconomic classes in the United States, or its impact, at this stage in the reduction of mortality, is relatively small compared to what could be achieved through reduction of the gap in levels of living and life styles associated with education, income, occupation, and geographic locale. If the United States is to demonstrate that she is indeed a land of equal opportunity, she must do considerably more to increase equality of opportunity on all fronts which affect the most significant index of effective equalitarianism--the ability to survive--duration of life itself.

These words were written in 1972 and referred to the authors' analyses of the cross-sectional 1960 Matched Records Study and of longitudinal census tract data for the city of Chicago. Socioeconomic differences in mortality were evident at both the individual and aggregate levels of analysis, no matter which indexes of socioeconomic level were employed. However, the longitudinal analysis of aggregated data for Chicago census tracts provided a finding which had special significance for the authors' conclusion regarding the improvement of social-economic conditions. They observed that between 1930 and 1940 there was a general convergence of socioeconomic differentials in the Chicago area, followed by a widening of these differentials between 1940 and 1960. As Kitagawa has more recently noted (1977), other research has also indicated a reversal of the older trend, i.e., now toward increasing socioeconomic differentials in mortality. For example, Lerner and Stutz (1976, 1977) have found widening differentials between 1960 and 1970 for Maryland and for the United States as a whole.

All of the studies which show a recent widening of socioeconomic differentials in the United States have been based solely on aggregate (or areal) data, employing "ecological" methods of analysis. Indeed, the largest part of research on mortality differentials has been based on aggregate data. Hannan and Burstein (1974) have noted that there generally will be a loss of efficiency for estimates from grouped observations. Moreover, using a structural equations perspective, they have shown that grouping of observations may result in biased estimates, depending on the nature of the causal relationships between the grouping criterion and the variables--both dependent and independent--in the model. Their analysis also emphasizes the possibility that grouping may have the effect of magnifying specification error in the micro-model of interest.

In view of these analytical considerations, we suggest that more attention should be given to the development of data systems which can provide individual socioeconomic characteristics in the analysis of trends in mortality. Accordingly, the purpose of this paper is to discuss important issues relating to the design of individual-level data systems with this goal.

Conceptualizing the variables.

One of the first concerns to be dealt with by anyone proposing an individual-level study of socioeconomic differentials in mortality is the problem of how to conceptualize the variables of interest. Generally, the resolution of this problem requires that we keep in mind how the parameter of common interest is calculated. We will usually want to obtain a rate for each socioeconomic group such that the weighted rates sum to the rate for the total population:

$$r_i = \left(\frac{\text{Deaths in class i during time period}}{\text{Population in class i during time period}}\right) xC$$

for each of K classes where each class is defined as a mutually exclusive subgroup of the total population such that

$$R = \text{rate for the total population} = \sum_{i=1}^{K} r_i p_i \text{ where}$$

p_i = proportion ith class is of the total population: $\sum_{i=1}^{K} p_i = 1.$

The problem which is immediately apparent even though the implications are not always realized is that a rate consists of a numerator and a denominator and that the classifications in the numerator and denominator should be identical. In forming an appropriate classification, the system must form classes which

1) are mutually exclusive and exhaustive of the population;
2) answer the question being asked;
3) make it possible to collect the data.

Creating mutually exclusive and exhaustive categories is a problem we always have to confront. The second and third considerations must always be faced as well, but because we are concerned here with mortality there are some extra problems which emerge. Among the socioeconomic characteristics of potential interest, some are fixed regardless of stage in the life cycle, some are stable (or at least relatively so) during adulthood, and some are subject to change over the entire life cycle. Examples of unchanging characteristics are sex, race, and ethnic group. Education and religion are characteristics that are relatively unlikely to change during adulthood, at least after age 25. Those characteristics changing throughout life clearly form the largest group, including age, marital status, size of family, living arrangements, quality of housing, employment status, labor force participation, occupation, income, assets, and residence.

From their analyses, Kitagawa and Hauser (1973) drew the conclusion that "education is probably the single most important indicator of socioeconomic status for mortality analysis." (p.179) Education was the measure they used to calculate excess deaths--the deaths which would not have occurred if the estimated age-specific death rates of white men (or women) who had completed at least one year of college had prevailed in each color-education subgroup of men (or women). It seems reasonable to infer that the usefulness of education as an indicator of socioeconomic status derives considerably from the stability of a person's educational level over adulthood.

If the characteristic of interest is one which changes over the life cycle, then the time reference is critical. For example, the question "Do mortality rates differ by income?" is deceptively simple and laden with traps for the unwary. The question must be clarified by stipulating a time frame. Specifically, we might refer to income at the time of death, but if we do so, we must be aware of the fact that two-thirds of the deaths in the United States are deaths after the 65th birthday when the majority of people are retired and probably have reduced incomes. For persons who die younger, it is possible that many such persons had to quit working because of the disability which led to death and consequently had unusually low incomes during the last year of life. Alternatively, we could be interested in maximum income earned during adulthood, or average annual income throughout adulthood. In the latter instances, it would be difficult to avoid expressing income in constant dollars. To study stress due to reduced income, the magnitude of the income reduction and the interval since it occurred would both be needed. To answer other types of questions, it might be necessary to obtain income of family during childhood, to supplement information on family background. In addition, it may be essential to distinguish between family income and individual income, because family size and relationships also change over time, and some people never do have any individual income. The answers to such questions will dictate the kinds of data one attempts to collect, and in turn the method of data collection. Viewed from the opposite direction, the limitations of the data collection system will modify the amount and type of data which can be collected, and the analytical design as well.

Data Collection Systems: A Typology.

It is useful to organize our discussion of issues related to the study of socioeconomic differentials in mortality by setting up a typology of possible mechanisms for collecting data on individual socioeconomic characteristics, as follows:

- Single systems: Numerator and denominator from the same source
  - Longitudinal
    - Population Registers
    - Prospective Studies designed for special purposes
  - Cross-sectional
    - Census of population
    - Interview surveys
      - Regular interview survey
      - Multiplicity survey
- Dual systems: Numerator and denominator from independent sources
  - Longitudinal
  - Cross-sectional
    - Record Matching
    - Follow-back surveys
      - Denominator from existing system
      - Denominator from special questions or systems

Single system longitudinal.

Longitudinal systems are those in which a cohort is defined by a characteristic or characteristics common to the group (born in a certain year, living in a specified area, members of a union) and the study group so defined is observed until the event of interest, in this case death, occurs. In a cohort study some of the relevant events may or may not have occurred at the time the cohort is defined but

death will not have occurred and the investigator must wait.

In theory, longitudinal systems are by far the best means of collecting data for differential mortality analysis. Data can be recorded on a continuing basis as people age so that there are no recall problems due to forgetfulness or bias because of later events.

The major disadvantages are due to the length of time involved. If data are needed to answer a current question, setting up a longitudinal data system now will not be useful. The cost of a longitudinal system is large as a staff has to be maintained over many years and the staff will change over the years as people involved in the original plan move on. Members of the cohort may be lost to observation unless very carefully followed and, if lost, must be traced to reduce bias.

Many of these disadvantages may be overcome if it is possible to tap into an existing system and utilize the data already collected.

In some countries there is a population register for the entire population which has to be updated each time an individual moves, changes jobs, or when other specified events occur.

The United States does not maintain a comprehensive population register. There are, however, a number of special registers which people stay on continuously. The Medical Follow-up Agency makes the medical experience of the general military-veteran population available and maintains a registry of 16,000 pairs of veteran twins as a subsidiary resource. There are disease registers, of which the cancer registers are probably best known. There are categorical program registers such as the Medicare recipients. There are registers maintained by some unions and professional organizations. For the most part these have not been utilized to study socioeconomic differentials in mortality and many of them in their present form cannot be used because the socioeconomic data are not recorded. It should be possible to add at least education to the data collected and thus increase their usefulness.

Prospective studies are designed for the specific purpose of following a cohort and recording observations about its members over a long period of time. They could be extremely useful for analysis of socioeconomic differentials if they were designed for that purpose, as the data are usually very carefully collected and recorded for the study participants.

There are two methodological problems with many of the prospective surveys now underway which make it impossible to draw inferences about socioeconomic differentials for the total population at risk. The first is that they are not probability samples. Many consist solely of white males who volunteer for the study and then remain participants on a voluntary basis. Some are restricted by the condition that the participants be healthy when the study began. The second problem is the well-known Hawthorne effect--the act of observing may change the characteristic being observed. The participants in a study usually receive some benefit from participation and the benefit is often early diagnosis or receipt of services which may affect the risk of death.

This is not to say that the prospective studies now underway are not useful or that a prospective study could not be designed to analyze socioeconomic differentials. The present studies are extremely useful for many purposes such as the study of physiological change. A study designed for socioeconomic analysis should be a probability sample of a defined population, must take into account the possible effects of observation on the participants, must have careful follow-up procedures for dropouts and analytical procedures for allowing for the dropouts, must be large enough to detect differences among the socioeconomic classes of interest, and must be well-funded over a period long enough for data collection and analysis.

Single system cross-sectional.

Cross-sectional studies are those in which data on the event of interest and the relevant variables all relate to the same point in time although the time reference may be extended through recall. When a single source is used to collect numerator and denominator data, the number of people who died and their characteristics must be obtained at the same time data on the population at risk is obtained. Collecting data on decedents in this fashion presents a number of methodological difficulties.

Any demographer knows that we have far better definitions of socioeconomic variables and far better data available for fertility than we do for mortality. One reason is the reality of funding; there has been far more funding for fertility research than for mortality research. A second, and more subtle reason, is that, given the paucity of information on either birth or death certificates, it is far easier to collect additional data on births than on deaths.

The usual method of collecting socioeconomic data is through a household interview census or survey. Such a survey works well for births, which are associated with family dissolution. It is possible through interviewing people in households to identify children by date of birth and collect the data of interest. In almost all cases the mother is living; in most cases the child is also. Contrast that with conducting household interviews to collect data on persons who died, say, within the year.

Two-thirds of the decedents in the United States are age 65 and over. In 1960, 4 percent of the population age 65 and over were residents of institutions, and 22 percent lived either

alone or with non-relatives. If there were no differential in death rates by living arrangements, that is, if death rates for people not living in families were the same as rates for people living in families, 22 percent of the elderly decedents would be missed on a census because there would be no surviving family member in the household to report for them and an additional 4 percent would be missed on an interview survey which did not cover residents of institutions.

However, death rates are not the same for elderly people in each type of living arrangement. In 1962-3, 23 percent of the elderly decedents were residents of institutions. Thirteen percent lived alone, and 4 percent lived with non-relatives. A question on the census would have missed 13-17 percent of the elderly decedents and a household survey would have missed 41 percent. Any analysis of death rates by socioeconomic status would be biased to the extent that socioeconomic status was associated with living arrangements. And that association does exist; people living alone or with non-relatives are poorer and less educated than those in families.

Among younger adults, the proportions living alone or in institutions are much lower but the differential death rates by living arrangement still exist. An additional problem is that when death occurs a household sometimes breaks up and reforms. The surviving member(s) move(s) in with someone else. There is no one in the original household left to interview. We do not have data on the extent of household reformation.

If a child dies, the household usually remains and data could be collected. Since deaths of children are rare events, the number of interviews required to yield a sufficient number of deaths for reliable estimates would be very large with consequent high cost.

One point that has been touched on needs to be stated explicitly. Age is important when considering the data needed and the best method of collecting it. Children are almost always living in families and their socioeconomic characteristics are those of the family. Adults under age 65 are usually living in families and the socioeconomic data of interest may be individual or family characteristics. Adults age 65 and over frequently are not living in families, the socioeconomic data of interest may be individual or family and may be current or from some time when they were eligible for employment, and household surveys do not include residents of institutions.

It is a shame that the household interview survey is not useful, as response rates for the continuing national surveys remain at approximately 95 percent. The effective ongoing data collection systems exist, but the disintegration of household of decedents and the fact that death is a rare event--on a population basis--preclude using this mechanism to collect data for the analysis of socioeconomic differentials in mortality.

A relatively new development in interview surveys is the multiplicity survey in which household respondents are asked to report not only for their own household members but also for a specified set of relatives (Sirken and Royston, 1970, 1973).

The advantages of a multiplicity survey are:

A. Smaller sampling errors than conventional survey;

B. Reduced response bias for decedents who lived alone at time of death, as a surviving relative in another household can report for them;

C. Can include institutional decedents.

The disadvantages of a multiplicity survey are:

A. Interviewer must collect the additional items;

B. Estimation and weighting procedures require carefully defined information;

   1. Household weight requires knowledge of the number of households containing persons eligible to report the death.

   2. Person weight requires knowledge of (a) the total number of persons eligible to report the death, and (b) the number of eligible persons living with the respondent. This is easier to collect because no knowledge is required of the location of other eligible persons.

No research has been done yet on whether the multiplicity approach will be useful for collecting socioeconomic data. Research to date has focused on how well the death itself has been reported and the basic demographic data.

Dual system longitudinal

It is possible to ascertain the fact of death from an independent source, usually the death certificate, and match that record with the records from a longitudinal data system or with record collected at some time in the past. This has in fact been done in epidemiological studies and has been especially useful in determining whether exposure to environmental conditions results in increased death risks.

Determining whether death has occurred and, if so, where (so that the death certificate can be located) is difficult and tedious. This has led to proposals for a National Death Index--a computerized register of all deaths occurring each year in the United States which could be used to ascertain whether an individual has died and in what State. Such a system would have all the problems inherent in any matching study but could greatly expand the potential

for socioeconomic analysis by providing the means for matching records from a census or survey with death records each succeeding year.

Dual system cross-sectional.

These systems, in which data on deaths are collected from the death registration system (or from surveys using it as a sampling frame) and data on the population are collected from another system, have been the only sources of National data on individual characteristics for the analysis of socioeconomic differentials in mortality.

The 1960 Matched Records Study is the prime example for the United States of using record linkage to provide nationwide information on socioeconomic differentials in mortality. The particular social and economic characteristics collected in the 1960 census, available on either Stage I records (complete enumeration) or Stage II records (25 percent sample), basically determined the operationalization of the social and economic differentials studied.

There was a total of 534,623 death certificates received by the National Center for Health Statistics in the period of May through August 1960. These deaths were taken as the universe in order to reduce the problems of matching death certificates to census schedules obtained in April 1960. To further reduce the cost of the manual search for matching records, half of the white decedents 65-74 were eliminated, and four-fifths of the white decedents over 74 were eliminated. This left a total of 340,033 death certificates to be matched to census schedules.

Next, the Bureau of the Census searched the complete enumeration schedules (Stage I) to link the 100 percent enumeration items with the death certificate information supplied by NCHS. Finally, those decedents matched with the first stage were matched to the second stage of the census, which contained much fuller socioeconomic information for a 25 percent sample of the population. As the Table 1 shows, 77 percent of the death certificates were matched to the 100 percent enumeration schedules. Of these 24 percent were matched with the sample enumeration schedules. Thus, about 18 percent of the decedents were available with full socioeconomic information. Among nonwhite decedents, the number of certificates not matched with Stage I schedules was about 50 percent higher than among white decedents. The potential for a racial bias is quite clear.

In order to estimate the "match bias" produced by failure to link certain decedents with Stage I schedules, the National Center for Health Statistics carried out a follow-back survey on a sample of decedents taken from the 340,033 decedents originally matched. It was intended that the results of this survey would enable researchers to make appropriate adjustments for bias, provided that the survey itself had minimal response bias. As it happened, although the census match rate was only 77 percent, the mail survey had a total response rate of 88 percent, and the personal interview follow-ups raised this to 94 percent (Table 2). When the response rates for unmatched white decedents were compared to those for the matched white decedents, it was found that the response rate varied between 87 and 93 percent (depending on the age group) for the unmatched group, and between 94 and 95 percent for the matched group. Thus there was very little relationship between match status of a decedent and the survey response for that individual. Moreover, the response rate for the matched group was somewhat higher than that for the total census schedule linkage to certificates. Kitagawa and Hauser concluded that

> The wide variations in nonmatch rates indicate that mortality differentials based on matched deaths alone would be subject to significant distortion and demonstrate the need for estimates of the social and economic characteristics of unmatched decedents.

As a result, the authors were forced to develop rather complex estimates of mortality ratios. Their decision not to calculate standard error estimates was also partly determined by the complexity of the ratio calculation procedure. And, of course, they were limited in analysis by the data available on the census. Essentially, socioeconomic status at time of death was the only information available for analysis.

The inability to match records for certain population subgroups is a reminder that there are serious biases using the census as a denominator for some forms of socioeconomic and mortality analysis, due to underenumeration on the census. Kitagawa and Hauser pointed out that differentials originally observed were reduced after they made corrections and that matching problems were especially serious for certain age, color, education, marital status and cause of death categories (TB, cirrhosis, accidents, and suicide). A recent paper by Rives demonstrates the effect of the 1970 census underenumeration on the life tables for the black population (Rives, 1977).

Interesting additional information on those problems comes from another matching study which immediately followed the Kitagawa and Hauser study. Records on psychiatric admissions in Louisiana and Maryland were matched with the census data for those two States. In this population, which was heavily weighted with poor people, black people, and people who had a high probability of being outside the mainstream of residing in nuclear households (the categories where underenumeration is a problem), the match rate was only 67 percent and the poorest rate for any diagnostic category was for alcoholics.

Finally, I'd like to note that you can only do a matching study when there is complete enumeration of the popualtion denominator to

match against. Heretofore, that has meant that matching studies of adult mortality could only be done every ten years--when there is a decennial census. The introduction of the quinquennial census will reduce this to every five years. The long intervals between censuses is a problem in areal studies as well.

Matching studies for infant mortality can be done at any time by matching against the birth certificate. The only disadvantage then is that one is limited to the information recorded on the two certificates. The birth certificate, unlike the death certificate, does have education on it. Because the birth and death occur so closely in time, are both recorded through the vital statistics registration system, and usually occur in the same State, problems of matching are vastly reduced. In comparison with matching problems on adult mortality, they are virtually eliminated.

In follow-back surveys, the numerator is a sample of the decedents and the denominator is from an independent data-collection system. The denominator can be from another set of records, a census, or from an independent population survey.

A national mortality survey was in operation at the National Center for Health Statistics on a continuing basis from 1961-1968 (in addition to the 1960 follow-back which supplemented the census match).

The procedure used in collecting the numerator data in this survey took advantage of the Current Mortality Sample, a 10 percent sample of deaths submitted by each State each month. This 10 percent sample was subsequently subsampled at a sampling rate of one out of 33, producing an overall rate of 1 out of 330 deaths registered in the United States. A mail survey then was the principal method of data collection. The primary source of information was the person who provided the funeral director with the personal information about the deceased for recording on the death certificate. The mailing address of the death record informant is usually reported on the death record but each primary source informant, attending physician, funeral director was asked to identify other persons who might be able to complete the questionnaire. Therefore, information was also collected from a secondary source if the primary source could not provide all of the requested information. There were also provisions for collecting missing information by other means; these included telephone and personal interviews which were carried out by the Bureau of the Census. Followup mailings were routinely sent to persons not responding, and other mailings were made to obtain complete and consistent information on the forms rejected as inadequate in a concurrent editing procedure. A poststratified ratio estimation procedure was used to make estimates.

The response rates for these surveys were about 90 percent; about 10 percent of the forms mailed to the informants either did not reach the informant or were not returned (Tables 3 and 4). The basic demographic information was available from the death certificate regardless of response and that information was used for imputation of the missing data.

The great advantage of collecting numerator data by this method (in addition to the high response rates and the provision for going to another source if the first one didn't know the information) is that questions asked on the follow-back survey can be matched precisely to the questions asked on the denominator data source. Wording and recall periods can be synchronized. Classification problems are minimized.

It is therefore possible to ask questions on a follow-back survey precisely as they are asked on the decennial census so that the concepts and categories are precisely the same without the necessity of matching. It is also possible to ask questions for infant deaths precisely as they are worded on the birth certificate.

However, by far the most flexible, and perhaps the most interesting, method of collecting denominator data is to have a concurrent survey especially designed to collect the data or to add special questions to an ongoing survey. Both approaches have been used.

In 1964-1966, the National Infant Mortality Survey--a follow-back based on infant death certificates--was in the field. During the same time period, the National Natality Survey--a follow-back based on birth certificates--was in the field to collect the denominator data. Response rates were high on both surveys (Tables 5 and 6). In June 1965 special questions were added to the Current Population Survey to serve as a denominator for the Natality Survey. The result was two sets of data on socioeconomic characteristics:

1964-66 National Infant Mortality Survey	Numerator
1964-66 National Natality Survey	Denominator
and	
1964-66 National Natality Survey	Numerator
June 1965 Current Population Survey	Denominator

Later, the 1966-68 Mortality Survey was devoted to questions on smoking. The same questions were asked on the Current Population Survey to provide precisely matched denominator data. Both surveys included questions on past history as well as current status.

In general, such an approach offers enormous flexibility for research. The matching of the

questions and the recall periods means that problems of recall, for example, are the same for both surveys. And you are not limited to the status at time of death; you can collect data about past history. The disadvantage is denominator data. Undercounting may still exist in an interview survey and residents of institutions are not included. However, they can be excluded from the numerator so that the universes are the same.

We would like to close with a few considerations other than response rates and matching--considerations which may overwhelm all statistical ones. Most important is cost. A follow-back survey is relatively inexpensive. The sampling frame is available through the continuous registration of deaths, sampling design is easy, estimation procedures are simple, and a mailed questionnaire is the interviewer. Data processing and analysis costs are the same as for other methods. A second consideration is the time it takes to complete a study. Data are more useful if they become available soon after the event of interest.

There is a great need for data for social epidemiology. Programs are being established, e.g., to pay for medical care for people in poverty and to provide services in areas where the median income is low, without enough data to help make intelligent decisions. One result of relying on area data has been that public services are located in areas where the median income is low, although there may be as many or more poor people living in areas with higher median incomes who do not have access to these services.

BIBLIOGRAPHY

Hannan, Michael T., and Leigh Burstein
1974 "Estimation from Grouped Observations," American Sociological Review, 39(June): 374-92.

Kitagawa, Evelyn M.
1977 "On Mortality." Presidential Address to the Population Association of America, April 22, St. Louis, Missouri.

Kitagawa, Evelyn M., and Philip M. Hauser
1973 Differential Mortality in the United States. Cambridge, Mass.: Harvard University Press.

Lerner, Monroe, and Richard N. Stutz
1976 "Socio-economic Differentials in Maryland, 1959-61 and 1969-71," Proceedings of the Social Statistics Section, 1976. Washington, D.C.: American Statistical Association.
1977 "Have We Narrowed the Gaps Between the Poor and the Non-Poor? Part II. Narrowing the Gaps, 1960-72: Mortality," to appear in Medical Care.

National Center for Health Statistics
1969 "Socioeconomic Characteristics of Deceased Persons," by Evelyn S. Mathis, Vital and Health Statistics, Series 22, No. 9. Washington, D.C.: Government Printing Office.
1972 "Infant Mortality Rates: Socioeconomic Factors, United States," by Brian MacMahon, Mary Grace Kovar and Jacob J. Feldman, Vital and Health Statistics, Series 22, No. 14, Washington, D.C.: Government Printing Office.

Rives, Norfleet W., Jr.
1977 "The Effect of Census Errors on Life Table Estimates of Black Mortality," American Journal of Public Health 67(9): 867-68.

Sirken, Monroe G., and Patricia Nellans Royston
1970 "Reasons Deaths are Missed in Household Surveys of Population Change," Proceedings of the Social Statistics Section, 1970. Washington, D. C.: American Statistical Association.
1973 "Underreporting of Births and Deaths in Household Surveys of Population Change," Proceedings of the Social Statistics Section, 1973. Washington, D.C.: American Statistical Association.

TABLE 1

Results of Matching 340,033 Death Records with 1960 Census Records, by Color and Sex: United States, May-August, 1960 Matched Records Study

Result of Census Match Operation	Total	White		All Other	
		Male	Female	Male	Female
Total Deaths in Match Operation	340,033	170,353	106,777	35,012	27,891
Deaths Matched with Stage I Census	262,966	133,921	85,484	23,836	19,725
Percent Matched	77.3	78.6	80.1	68.1	70.7

Source: Kitagawa and Hauser, Differential Mortality in the United States, 1973, p. 187.

TABLE 2

Response to NCHS 1960 Follow-back Survey, for 8,121 Decedents 25 years of age and over, by Color and Sex and Whether or not Matched on Stage I Census Record

Response to NCHS Survey	Total	White		All Other	
		Male	Female	Male	Female
Total Decedents in Survey	8,121	4,199	2,936	542	444
Responded to Survey	7,580	3,936	2,762	483	399
Percent Responded	93.3	93.7	94.1	89.1	89.9
Matched with Census	6,481	3,384	2,379	392	326
Responded to Survey	6,108	3,198	2,257	355	298
Percent Responded	94.2	94.5	94.9	90.6	91.4
Unmatched with Census	1,640	815	557	150	118
Responded to Survey	1,472	738	505	128	101
Percent Responded	89.8	90.6	90.7	85.3	85.6

Source: Kitagawa and Hauser, Differential Mortality in the United States, 1973, pp. 189-190.

TABLE 3

Number of Sample Cases and Percent for Which Response was Received, by Age, Color, and Sex of Decedents: 1962-65 National Mortality Surveys

Age, color, and sex of Decedents	Number	Percent with Responses
All decedents	22,948	90.5
Age in years		
Under 1	2,392	85.1
1-14	423	84.2
15-24	362	88.1
25-44	1,314	89.1
45-54	1,907	88.5
55-64	3,406	88.7
65-74	5,274	91.8
75 and over	7,870	93.1
Color		
White	19,982	91.3
All other	2,966	85.3
Sex		
Male	13,053	90.3
Female	9,895	90.8

Source: Unpublished data from the Division of Vital Statistics, National Center for Health Statistics.

TABLE 4

Number and Percent Responding to Informant Questionnaire in the National Mortality Survey, 1966-68

Year of Survey	Number of Decedents in Sample	Percent with Responses
Total	19,526	92.3
1966	6,391	93.6
1967	6,370	94.8
1968	6,765	88.6

Source: Unpublished data from the Division of Vital Statistics, National Center for Health Statistics

TABLE 5

Number and Percent Responding by Selected Characteristics of Mothers in the National Natality Survey, 1964-66

Characteristics of Mother	Number in Survey	Percent Responding
Total	10,395	88.8
Age		
Under 20 years	1,466	82.5
20-24 years	3,698	88.7
25-29 years	2,617	90.7
30-34 years	1,562	90.7
35 years and over	1,052	90.5
Color		
White	9,096	89.5
All other	1,299	84.0
Live-birth order		
First	3,009	88.7
Second	2,596	89.4
Third	1,852	89.4
Fourth	1,208	89.1
Fifth or higher	1,730	87.2
Region of residence		
Northeast	2,445	92.8
North Central	2,968	91.4
South	3,246	87.1
West	1,736	82.0
Metropolitan Status		
Inside SMSA	6,682	90.4
Outside SMSA	3,713	85.9

Source: National Center for Health Statistics, Vital and Health Statistics, Series 22, No. 14

TABLE 6

Number and Percent Responding to Informant Questionnaire by Selected Characteristics of Deceased Legitimate Infants in the National Infant Mortality Survey, 1964-65

Characteristics of Deceased Infants	Total Number of Legitimate Infants	Percent with Response
Total	1,497	87.9
Race		
White	1,164	88.7
Black	302	86.4
Other	31	71.0
Region		
Northeast	302	90.7
North Central	439	89.5
South	515	89.3
West	241	76.4
Metropolitan status		
Metropolitan	907	88.9
Nonmetropolitan	590	86.4
Age at death		
Under 1 day	613	88.3
1-6 days	361	89.5
7-27 days	105	85.7
28 days-5 months	293	87.4
6-11 months	125	84.8

Source: National Center for Health Statistics, Vital and Health Statistics, Series 22, No. 14.

RESEARCH ON THE RELATIONSHIP BETWEEN SOCIOECONOMIC STATUS AND MORTALITY IN THE UNITED STATES: 1960-1975

Edward G. Stockwell, Jerry W. Wicks, and Donald J. Adamchak
Bowling Green State University

The present paper has two specific aims: first to summarize, as succinctly as possible, the present state of our knowledge concerning the nature of this differential in the United States today; and second to suggest the kinds of research that still needs to be done to increase our knowledge of this differential so that we may take further steps to eliminate it. The paper is divided into two main sections: the first will consider the relationship between socioeconomic status and mortality in general, whereas the second will look at the situation as it pertains to infant mortality. I make this distinction for three reasons: (1) infant mortality has long been recognized as the most sensitive mortality indicator of group differences in social and economic well-being; (2) it is the aspect of mortality on which my own research has concentrated and with which I am most familiar; and (3) perhaps most important, very different kinds of research are needed for a more adequate understanding of the different "causes" of the infant mortality/socioeconomic status relationship as opposed to those characterizing total mortality and socioeconomic status.

SOCIOECONOMIC STATUS AND MORTALITY IN THE UNITED STATES

It has been over a decade since anyone has presented a review of the research findings on this topic. At that time (early 1960's) two reviews were published which seemed to suggest that there was some basis for optimism with regard to the future course of the socioeconomic mortality differential. In the first of these (Stockwell, 1961) it was noted that although most of the studies that had been carried out in the post-World War II era revealed the existence of a fairly pronounced inverse relationship between mortality rates and socioeconomic status, there nevertheless seemed to be emerging differences as to the magnitude of the differential, and as to whether or not it was narrowing. Based on a review of several studies done during the 1950 decade, as well as on the results of some of my own research (Stockwell, 1963), it was concluded that both the extent of the socioeconomic differential and the nature of its trend depended on such things as the area under investigation, the particular variables used to measure socioeconomic status, and the nature of the methodological procedures followed. Further, the very fact that what had previously been a consistent and pronounced inverse association had become so variable was sufficient to encourage speculation about an emerging trend toward a closing of the socioeconomic status mortality gap.

In the second review (Antonovsky, 1967), somewhat similar conclusions were reached. Although it was emphasized that a socioeconomic differential still existed, there was clearly a trend toward a blurring of the traditional pattern. Specifically, it was noted that the differentials then observed were pretty much limited to a difference between the lowest class and all others. That is, what had once been a fairly smooth inverse gradient across several socioeconomic class levels was now one in which similar low death rates characterized all the upper and middle class groupings, with a much higher death rate prevailing in the lowest group. This blurring of the traditional inverse relationship was explained in terms of the continuation of the historical decline of mortality in our society. That is, it was suggested that when mortality levels are extremely high or extremely low (i.e., at the two extremes when men either have very little control over their life chances or when they have achieved a great deal of success in controlling mortality), social class differences will be small; and further that it is during the transitional phase from high to low death rates, when the fruits of health progress filter slowly down from the richer to the poorer classes, that the socioeconomic differential is most apparent. This being the case it would suggest the hypothesis that as the overall death rate of a population was lowered further the remaining class differences would decline. Although the lowest socioeconomic groups were still characterized by a notable mortality disadvantage, the fact that the mortality levels of all other classes had blurred clearly suggested that this differential was not inevitable and that it could become even more blunted with further advances in the control or mortality.

Research Since 1960: Basically we can distinguish between two kinds of studies that have examined the relationship between socioeconomic status and mortality: those which have collected data for individuals, and those which have been based on data for ecological units -- particularly census tracts. By far most of the research on this topic has been of the second type (very likely reflecting the cost differences in carrying out these two kinds of studies and, related, the relative absence of funding to support social research on mortality). Nevertheless, at least two noteworthy efforts of the first type are represented by (1) the National Mortality Surveys and birth/death linkage studies done by the National Center for Health Statistics during the 1960's, and by (2) the fairly detailed census-death certificate matching study reported by Kitagawa and Hauser (1973). While such studies using individual data are necessary for a full understanding of the nature and causes of the socioeconomic mortality differential, the fact that there have been so few of them (especially the lack of comparable studies over time) seriously limits the kind of conclusions that can be drawn from them.

Turning now to a brief consideration of the more common census tract based studies of socioeconomic status and mortality, the most overriding conclusion that seems to be warranted is that, contrary to the earlier optimistic speculations, there has been little if any change in the situation since the 1950's. Recent studies, in fact,

have revealed that a strong socioeconomic mortality differential characterizes cities as diverse in size and characteristics and as widely separated in space as Lexington, Kentucky (Quinney, 1965), Columbus, Ohio (Schwirian and Lagreca, 1971), Chicago, Illinois (Kitagawa and Hauser, 1973), Hartford, Connecticut (Nagi and Stockwell, 1973), and Phoenix and Tucson, Arizona (See Table 1). Beyond noting that it still exists, however, one has to conclude that the precise nature of this differential is still inadequately understood. To illustrate, there is disagreement as to whether it characterizes all segments of the population. In the study of Lexington, for example (Quinney, 1965), in which three separate measures of socioeconomic status and a combined index were used, very little association was found between socioeconomic status and mortality for the young adult group (ages 20-39). This observation conflicts with both the findings of a number of earlier studies (Antonovsky, 1967) and with more recent data (see Table 2) which suggest that the socioeconomic differential is very pronounced among the early adult ages, particularly ages 30-39. Similarly, although the same Lexington study revealed a positive association between socioeconomic status and mortality for nonwhites, data for Chicago in 1960 and for both Phoenix and Tucson in 1970 indicate that the inverse differential is just as pronounced for nonwhites as it is for whites (see Table 3).

The particular index of socioeconomic status used does not seem to effect the existence of the relationship, but there is some variation as to its magnitude, and such variation could be significant for the kind of conclusions drawn. Most of these city socioeconomic areas are based on median family income (Chicago, Lexington, Phoenix and Tucson), and where several indices were used (Quinney, 1965), the highest correlation between socioeconomic status and mortality was found to characterize the income variable. In Columbus, Ohio, however, Schwirian and Lagreca (1971) found that housing conditions (percent of dwelling units in sound condition) were much more highly correlated with mortality rates than was median family income.

To cite one other illustration, the data presented in Table 1 would suggest that the nature of the socioeconomic differential by sex is also unstable. As would be expected, female death rates are everywhere lower than corresponding male rates; however, the relative difference between the lowest and highest economic areas is notably greater for females at every year in Chicago; but it is substantially more pronounced for males in both Phoenix and Tucson. Finally, with respect to the earlier postulated blurring of class lines above the lowest group, the data presented in Table 1 would suggest that this may be the trend for females, but that such a blurring has not characterized males to the same extent -- particularly in the two Arizona cities.

What these isolated findings from a few selected studies indicate, then, is that we are still pretty much where we were at the start of the 1960 decade. We know without question that a low socioeconomic status is associated with a higher than average death rate, but when it comes to making more specific conclusions there is still a good deal of variation from one area to another, from one population subgroup to another, and from one measure of socioeconomic status to another.

What is more important, however, is that we have not made much progress in explaining what it is about a low socioeconomic status that results in the higher death rates; and the unfortunate corrolary is the already noted fact that we have not made any real progress in eliminating or reducing this differential. Beyond some noteworthy attempts to isolate the socioeconomic status component that contributes most to the differences in mortality -- for example, the specification by Schwirian and Lagreca (1971: 585-587) that the effect of status on mortality operates through the housing variable, and likely reflects such concomitants of poor housing conditions as inadequate lighting, heating and sanitation, as well as the higher incidence of certain social problems like alcoholism, broken homes and drug addiction -- ...beyond such efforts there has been a lot of speculating and hypothesizing, but very little real research, relating to the influence of such things as genetic inheritance (Quinney, 1965), and to differences in health care knowledge and access to good medical care, especially preventive care (Antonovsky, 1967: 67). And the need for research with respect to these kinds of factors is especially important today as the influence of infectious diseases has declined and as the chronic diseases, particularly heart disease, have assumed a greater responsibility for the pronounced mortality disadvantage characterizing the lowest socioeconomic groups in our society (Quinney, 1965; Nagi and Stockwell, 1973).

Before we can suggest realistic remedial programs we need to know a lot more about the problem with which we are confronted. Part of the problem to date stems from the past heavy reliance on the use of ecological data to study the relationship between socioeconomic status and mortality, and this in turn is at least partly due to a deficiency of monies available for social epidemiological research on mortality. In order to isolate the specific factors involved and to arrive at a more adequate understanding of the underlying causes of the socioeconomic mortality differential (for the general population and for particular ethnic subgroups within it) we need both the extensive surveys and the intensive case studies of the kind that we have so long had with respect to fertility.

INFANT MORTALITY AND SOCIOECONOMIC STATUS

Although the infant mortality rate has long been recognized as an extremely sensitive index of differences in the levels of social and economic well-being characterizing various geographic areas or population subgroups (Newsholme, 1910; Woodbury, 1925), and although numerous studies suggest that infant mortality continues to be highly sensitive to socioeconomic differences on an international level (Ekanem, 1972; Stockwell, 1960 and 1966; Stockwell and Hutchinson, 1975), a number of studies published in the early 1960's raised questions concerning the precise status of this traditionally inverse relationship within an

advanced, relatively low mortality country such as the United States (Donabedian, et al., 1965; Stockwell, 1962; Willie, 1959). These questions have arisen largely as a consequence of the marked declines in infant mortality rates in modern, industrial societies (Chase, 1967), particularly the declines in the postneonatal component of infant mortality. These latter studies suggested that in countries where infant mortality was low, and where the major proportion of infant deaths occur in the neonatal period and are attributed to endogeneous causes (e.g., immaturity, birth injury, congenital malformations, postnatal asphxia), the traditional negative correlation between infant mortality and socioeconomic status would be blunted. On the other hand, for those few deaths that do take place between the ages of one month and one year, where the major causes of death are further removed from the physiological processes of gestation and birth, mortality levels would continue to exhibit an inverse relationship to socioeconomic status. At least one of these studies went even further and suggested that continued progress in the public health and medical professions could, by contributing to still greater reductions in the proportion of infant deaths occurring in the postneonatal period, blunt the traditional association even further -- and perhaps even eliminate it (Stockwell, 1962).

What has happened to the traditional inverse relationship between infant mortality and socioeconomic status? Once again, an examination of the findings and conclusions of more recent studies does not yield a definitive answer. To illustrate, although a longitudinal study of infant mortality in the Chicago area showed a marked narrowing of the socioeconomic differential between 1930 and 1960 (Kitagawa and Hauser, 1973: 66-67), other data for New York City (National Academy of Science, 1973), Toledo, Ohio (Adamchak, et al., 1976), San Antonio, Texas (Markides and Barnes, 1977), the state of Ohio (Stockwell and Laidlaw, 1977), and for the nation as a whole (Kitagawa and Hauser, 1973: 28-29; MacMahon et al., 1972), suggest that the traditional relationship is just as pronounced as ever. Furthermore, still other research has noted that the inverse relationship is also characteristic of the neonatal component of infant mortality, not only in the United States (Shapiro, et al., 1968; Brooks, 1975; Shin, 1975; Adamchak and Stockwell, 1977; Stockwell and Laidlaw, 1977) but also in other industrialized low mortality countries (Douglas, 1966; de Haas-Posthuma and de Haas, 1968; Hirst et al., 1968).

The preceding discussion clearly reveals a lack of consistency among conclusions pertaining to the relationship between infant mortality and socioeconomic status. Some of the confusion, of course, reflects the fact that the studies cited are based on a variety of units of analysis (matched records, census tracts, states) and have used different measures of socioeconomic status (mother's education, father's occupation, family income). It may also reflect real differences among the population groups studied (i.e., the earlier studies that questioned the traditional relationship were all carried out in the urban northeastern region of the United States, and those national data that are available indicate the relationship is least pronounced in the northeast) (MacMahon, et. al., 1972:5). Further, those studies that have talked about the changing pattern of this relationship have generally been cross-sectional in nature, inferring change by comparing their findings with those of earlier studies (most of which were carried out in different areas and based on different methodologies). In short, this is clearly a topic where additional research is sorely needed.

Preliminary Results of an On-going Study: Staff members of the Department of Sociology at Bowling Green State University are presently engaged in a fairly broad study of the relationship between socioeconomic status and mortality, one phase of which is a longitudinal study of the trend with respect to infant mortality within the major metropolitan areas of Ohio. Data from this study are presently available for the city of Toledo, for 1950 and 1970, and some preliminary results of our analysis are included here in Table 4. The zero order correlation coefficients presented here for 1950 would clearly tend to support the conclusions of earlier studies that postulated a blunting of the traditional infant mortality/socioeconomic status association -- a blunting that seemed to be explainable in terms of the lack of any significant relationship between socioeconomic status and the neonatal component of infant mortality. However, it is equally clear that the projected further blunting of the overall association has not been realized. In fact, the relationship for total infant mortality is more pronounced in 1970 than it was in 1950 for all three socioeconomic indicators.

Further examination of these data indicates that the relationship with respect to postneonatal mortality has declined (although not significantly) for two of the three socioeconomic indexes, whereas the relationship with respect to neonatal mortality has increased significantly for all socioeconomic measures. The net effect of these two trends has been to create a situation in 1970 where, with the exception of the income measure, the strength of the mortality/socioeconomic status relationship is greater for the neonatal death rate than it is for the postneonatal. (The difference between neonatal and postneonatal with respect to the income measure is so small it can be regarded as inconsequential).

These findings, are consistent with those of at least one other recent study (Brooks, 1975), and are clearly not in line with what would have been expected on the basis of research done 10 to 15 years ago; and they give rise to two important questions:

(1) What has caused the overall relationship between infant mortality and socioeconomic status to increase?
(2) What has caused the emergence of the neonatal component as the major contributor to the overall relationship?

With respect to the first question, one factor may be the nature of recent migration patterns and the changing composition of the urban population -- particularly the increase in the proportion of Blacks among the infant deaths in Toledo (from 17 percent in 1950 to 37 percent in 1970). Since Blacks are overrepresented in the

poorest socioeconomic areas, and since the traditionally more sensitive postneonatal mortality accounts for a larger proportion of Black infant deaths (Kleinman, et al., 1976), an increasing proportion of Blacks in the study population may be contributing to the stronger association during the more recent period. This is a question that is currently being explored further.

The second question poses greater difficulties. On the one hand, it may be that the increase in the magnitude of the neonatal/socioeconomic relationship is also explainable, at least in part, by the increasing proportion of Blacks in the study population. If, for example, the neonatal/socioeconomic relationship were to be more pronounced for Blacks than for the white population, then the sizable increase in the Black fraction could very easily be "overpowering" the lesser relationship among whites in the more recent period, (e.g., low birth weight, a major contributor to infant death, is about twice as prevalent among Blacks). On the other hand, the changing patterns of the association between infant mortality and socioeconomic status may reflect some as yet undetected changes with respect to the role of particular causes of death. For example, our data indicate that for Toledo, in direct contrast to the national trend, there has been an increase in the proportion of infant deaths occurring in the postneonatal period. Why this should be the case is still unclear to us, and is one of the key questions still under investigation. (Again there is probably an association with the changing composition of the population in many of our urban centers).

Another explanation that has been suggested is that the exogenous causes of death more commonly associated with postneonatal mortality are now contributing to neonatal mortality. A specific factor here could be the nutritional status of the mother's diet during pregnancy, as it is known that lower socioeconomic groups have a nutritionally poor diet relative to that of the general population (Belli, 1971; Chabot et al., 1975), and this could be a factor contributing to the higher incidence of low birth weight babies among low socioeconomic groups.

In conclusion we would emphasize that we still do not have a definitive answer to the general question "What is happening to the relationship between infant mortality and socioeconomic status?" This evidence from our very preliminary work to date suggests that there has indeed been a major shift away from what appeared, 10 to 15 years ago, to be a contracting association back to a clear-cut and very pronounced negative relationship. The explanation of this changing pattern is far from clear, however; and it is this that will be the major focus of our continuing research on this topic. It is very doubtful, however, if our research will provide answers to all of the relevant questions. On the one hand, data on such things as the quality of prenatal care, diet, and infant care knowledge and practices are not available in ecological analyses such as ours. On the other hand, a lot of relevant data that are available on the birth record -- parity, length of gestation, birth weight -- are not readily accessible to us on an individual basis. As with mortality in general, such ecological analyses are clearly insufficient. Birth-death record link studies are a positive step in the right direction (Armstrong, 1972; Chase, 1972), but they too are insufficient (e.g., they do not get at maternal habits and life style). What we really need in order to increase our knowledge of the relative effect of the specific factors responsible for higher infant death rates among the lower socioeconomic groups is extensive studies that look at infants who die at various ages and those who survive the first year of life in terms of a wide variety of individual and family life style characteristics.

* * * * * * *

Table 1. -- Age-standardized average annual death rates per 1,000 population for five social rank areas, white population by sex, for various cities and dates.

City and Year			Socioeconomic					Ratio
			I (High)	II	III	IV	V (Low)	V:1
Chicago,	1930	M	11.6	12.4	13.6	15.4	18.8	1.62
		F	6.6	7.2	8.4	9.9	13.2	2.00
Chicago,	1940	M	11.0	10.8	11.5	13.4	16.6	1.51
		F	5.8	5.6	6.3	7.8	10.4	1.79
Chicago,	1950	M	8.7	9.4	9.7	11.6	14.6	1.68
		F	4.2	4.9	5.1	6.4	8.6	2.05
Chicago,	1960	M	9.6	9.2	10.1	11.3	16.0	1.67
			4.7	4.5	5.2	6.0	8.6	1.83
Houston,	1950	M	7.5	7.9	9.1	11.1	9.9	1.32
		F	5.4	5.3	5.6	7.1	7.5	1.39
Providence,	1950	M	10.8	11.8	11.2	12.7	14.0	1.30
			7.3	7.6	8.9	9.4	10.4	1.42

Table 1 Con't.

City and Year		Socioeconomic					Ratio
		I (High)	II	III	IV	V (Low)	V:I
Hartford, 1950	M	9.3	10.3	11.2	11.8	12.5	1.34
	F	6.6	7.5	7.5	8.2	8.3	1.26
Phoenix, 1970	M	9.8	10.9	11.5	13.4	18.2	1.86
	F	6.4	6.6	6.4	7.2	8.9	1.39
Tucson, 1970	M	8.8	9.9	9.5	11.5	15.1	1.72
	F	6.3	6.3	5.0	6.6	7.9	1.25

SOURCES: Chicago data (Kitagawa and Hauser, 1973, p. 53); Hartford, Providence, and Houston (Antonovsky, 1967, p. 54); Phoenix and Tucson (calculated by authors from data supplied by the Arizona Department of Health).

* * * * *

Table 2. -- Age-specific white death rates, by sex, for highest and lowest social rank areas in Phoenix, 1970

Age	High SES	Low SES	Ratio Low:High	High SES	Low SES	Ratio Low:High
	White Males			WHITE FEMALES		
0-1	11.8	21.3	1.81	13.2	11.9	.90
1-9	0.8	1.2	1.50	0.3	1.3	4.33
10-19	0.8	1.3	1.63	0.4	0.7	1.75
20-29	2.7	4.0	1.48	0.7	1.2	1.71
30-39	1.6	7.6	4.75	0.8	3.8	4.75
40-49	3.2	16.3	5.09	2.6	5.9	2.27
50-59	9.6	30.7	3.20	5.3	11.6	2.19
60-69	36.6	56.9	1.55	13.1	21.4	1.63
70+	77.3	102.4	1.32	61.2	64.9	1.06

SOURCE: Calculated by authors from data supplied by the Arizona Department of Health.

Table 3. -- Age-standardized death rates of nonwhites, for high and low social rank areas, for various cities and dates.

City and Year	High SES	Low SES	Ratio Low:High
Chicago, 1960			
Male	9.8	16.7	1.70
Female	8.1	11.6	1.42
Phoenix, 1970	7.3	12.0	1.64
Tucson, 1970	5.7	9.6	1.68

SOURCES: Chicago (Kitagawa and Hauser, 1973, pp. 54-55); Phoenix and Tucson (calculated by authors from data supplied by the Arizona Department of Health).

* * * * *

Table 4. -- Zero order correlation coefficients between infant mortality and three measures of socioeconomic status: Toledo, Ohio, 1950 and 1970

Infant mortality component and socioeconomic measures[1]	Correlation coefficients 1950	1970	Difference, 1950-1970 Absolute difference	Level of significance
Total infant mortality				
Education	-.297*	-.500***	+.203	.11%
Occupation	-.288*	-.549***	+.211	.10
Income	-.267*	-.667***	+.400	.004
Neonatal				
Education	-.113	-.430**	+.317	.04
Occupation	-.120	-.451***	+.331	.03
Income	-.119	-.528***	+.409	.01
Postneonatal				
Education	-.435**	-.356**	-.079	.32
Occupation	-.402**	-.328*	-.074	.33
Income	-.357**	-.530***	+.173	.14

* = Significant at .05 percent.
** = Significant at .01 percent.
*** = Significant at .001 percent.

1. Education, the median number of school years completed by persons age 25 years and over; occupation, the percent of the employed population engaged in white collar occupations; and income, median income of families and unrelated individuals. The unit of analysis is the census tract of mother's residence.

* * * * *

References and Selected Bibliography

Adamchak, Donald J., Robert E. Siegel and Edward G. Stockwell. "Infant Mortality and Socioeconomic Status." Ohio's Health, 28 (July-August, 1976): 8-9.

Adamchak, Donald J. and Edward G. Stockwell. "Trends in the Relationship Between Infant Mortality and Socioeconomic Status: 1950-1970." Paper presented at the annual meeting of the North Central Sociological Association (Pittsburgh, May, 1977).

Antonovsky, Aaron. "Social Class, Life Expectancy and Overall Mortality." Milbank Memorial Fund Quarterly, 45 (April, 1967): 31-73.

Armstrong, Robert J. "A Study of Infant Mortality from Linked Records by Birth Weight, Period of Gestation and Other Variables." National Center for Health Statistics, Vital and Health Statistics, 20:12 (May, 1972).

Belli, Pedro. "The Economic Implications of Malnutrition: The Dismal Science Revisited." Economic Development and Cultural Change, 21 (October, 1971): 1-23.

Bedger, Jean E., Abraham Gelperin and Eveline E. Jacobs. "Socioeconomic Characteristics in Relation to Maternal and Child Health." Public Health Reports, 81 (September, 1966): 829-833.

Bendor, Daniel E. et al. "Factors Affecting Postneonatal Mortality." HSMHA Health Reports, 86 (May, 1971): 482-486.

Bouvier, Leon F. and Jean van der Tak. "Infant Mortality: Progress and Problems." Population Bulletin, 31 (April, 1976).

Brooks, Charles H. "The Changing Relationship Between Socioeconomic Status and Infant Mortality: An Analysis of State Characteristics." Journal of Health and Social Behavior, 16:3 (1975): 291-303.

Chabot, Marion J., Joseph Garfinkel and Margaret W. Pratt. "Urbanization and Differentials in White and Nonwhite Infant Mortality." Pediatrics, 56:5 (1975): 771-781.

Chase, Helen C. "International Comparison of Perinatal and Infant Mortality: The United States and Six West European Countries." National Center for Health Statistics, Vital and Health Statistics, 3:6 (March, 1967).

_____. "A Study of Infant Mortality from Linked Records: Comparison of Neonatal Mortality from Two Cohort Studies." National Center for Health Statistics, Vital and Health Statistics, 20:13 (June, 1972).

Committee on Maternal and Child Care. "Reducing Infant Mortality." Journal of the American Medical Association, 193 (July, 1965): 310-319.

de Haas-Posthuma, J. H. and J. H. de Haas. "Infant Loss in the Netherlands." National Center for Health Statistics, Vital and Health Statistics, 3:11 (August, 1968).

Donabedian, Avedis, Leonard S. Rosenfeld, and Edward M. Southern. "Infant Mortality and Socioeconomic Status in a Metropolitan Community." Public Health Reports, 80 (December, 1965): 1083-1094.

Douglas, Charlotte, A. "Infant and Perinatal Mortality in Scotland." National Center for Health Statistics, Vital and Health Statistics, 5:3 (November, 1966).

Ekanem, Ita I. "A Further Note on the Relation Between Economic Development and Fertility." Demography, 9 (August, 1972): 383-398.

Falkner, Frank, ed. Key Issues in Infant Mortality, National Institute of Child Health and Human Development, 1969.

Hirst, Katherine M., Neville R. Butler, and M. J. R. Dawkins. "Infant and Perinatal Mortality in England and Wales." National Center for Health Statistics, Vital and Health Statistics, 3:12 (November, 1968).

Hunt, Eleanor P. "Lags in Reducing Infant Mortality." Welfare in Review, 2 (April, 1964): 1-14.

_____. "Infant Mortality and Poverty Areas." Welfare in Review, 5 (August-September, 1967): 1-12.

Hunt, Eleanor P. and Earl E. Huyck. "Mortality of White and Nonwhite Infants in Major U.S. Cities." HEW Indicators (January, 1966): 23-41.

Jiobu, Robert M. "Urban Determinants of Racial Differentiation in Infant Mortality." Demography, 9 (November, 1972): 603-615.

Kitagawa, Evelyn M. and Philip M. Hauser. "Education Differentials in Mortality by Cause of Death: United States, 1960." Demography, 5 (February, 1968): 318-353.

_____. Differential Mortality in the United States: A Study in Socioeconomic Epidemiology. Harvard University Press, 1973.

Kleinman, Joel C., Jacob J. Feldman and Robert H. Mugge, "Geographic Variations in Infant Mortality." Public Health Reports, 91 (September-October, 1976): 423-432.

MacMahon, Brian, Mary Grace Kovar, and Jacob J. Feldman. "Infant Mortality Rates: Socioeconomic Factors." National Center for Health Statistics, Vital and Health Statistics, 22:14 (March, 1972).

Markides, Kyriakos S., and Donna Barnes. "A Methodological Note on the Relationship Between Infant Mortality and Socioeconomic Status with Evidence from San Antonio, Texas," Social Biology, 24 (Spring, 1977): 38-44.

Mathis, Evelyn S. "Socioeconomic Characteristics of Deceased Persons." National Center for Health Statistics, Vital and Health Statistics, 22:9 (February, 1969).

Moriyama, Iwao M. "Present Status of Infant Mortality Problem in the United States." American Journal of Public Health, 56 (April, 1966): 523-625.

Morris, Naomi M., J. Richard Udry, and Charles L. Chase. "Shifting Age-Parity Distribution of Births and the Decrease in Infant Mortality." American Journal of Public Health, 65 (April, 1975): 359-362.

Nagi, Mostafa H. and Edward G. Stockwell. "Socioeconomic Differentials in Mortality by Cause of Death." Health Services Reports, 88 (May, 1973): 449-456.

National Academy of Sciences. Infant Death: An Analysis by Maternal Risk and Health Care. Institute of Medicine, 1973.

Newman, John F. and William L. Graves. "Neonatal Mortality and Socioeconomic Status in a Metropolitan County." Sociological Symposium, 8 (Spring, 1972): 37-49.

Newsholme, Arthur. Thirty-ninth Annual Report of the Local Government Board, Report C d 5312. Darling and Son, Ltd., 1910.

Norris, Frank D. and Paul W. Shipley. "A Closer Look at Race Differentials in California's Infant Mortality, 1965-67." HSMHA Health Reports, 86 (September, 1971): 810-814.

Omran, Abdel R. "Epidemiologic Transition in the U.S." Population Bulletin, 32 (May, 1977).

Patno, Mary Ellen. "Mortality and Economic Level in an Urban Area." Public Health Reports, 75 (September, 1960): 841-851.

Quinney, Richard. "Mortality Differentials in a Metropolitan Area." Social Forces, 43 (December, 1965): 222-230.

Roberts, Robert E. and Cornelius Askew, Jr. "A Consideration of Mortality in Three Subcultures." Health Services Reports, 87 (March, 1972): 262-270.

Shah, Farida K. and Hellen Abbey. "Effects of Some Factors on Neonatal and Postneonatal Mortality." Milbank Memorial Fund Quarterly, 49 (January, 1971): 33-57.

Shapiro, Sam and Iwao M. Moriyama. "International Trends in Infant Mortality and Their Implications for the United States." American Journal of Public Health, 53 (May, 1963): 747-760.

Shapiro, Sam, Edward R. Schlesinger, and Robert E. L. Nesbitt, Jr. Infant, Perinatal, Maternal, and Childhood Mortality in the United States. Harvard University Press, 1968.

Shin, Eui Hang. "Black-White Differentials in Infant Mortality in the South, 1940-1970." Demography, 12 (February, 1975): 1-19.

_____. "Economic and Social Correlates of Infant Mortality: A Cross-sectional and Longitudinal Analysis of 63 Selected Countries." Social Biology, 22 (Winter, 1975): 315-325.

Schwirian, Kent and Anthony J. Lagreca. "An Ecological Analysis of Urban Mortality Rates." Social Science Quarterly, 52:3 (1971): 574-587.

Steahr, Thomas, E. "Mortality by Socioeconomic Levels." Connecticut Medicine, 40 (August, 1976): 553-555.

Stockwell, Edward G. "The Measurement of Economic Development." Economic Development and Cultural Change, 8 (July, 1960): 419-432.

_____. "Socioeconomic Status and Mortality in the United States." Public Health Reports, 76 (December, 1961): 1081-1086.

_____. "Infant Mortality and Socioeconomic Status: A Changing Relationship." Milbank Memoral Fund Quarterly, 40 (January, 1962): 101-111.

_____. "A Critical Examination of the Relationship Between Socioeconomic Status and Mortality." American Journal of Public Health, 53 (June, 1963): 956-964.

_____. "Infant Mortality and Socioeconomic Status." Connecticut Health Bulletin, 79 (November, 1965): 259-262.

_____. "Some Demographic Correlates of Economic Development." Rural Sociology, 31 (June, 1966): 216-224.

Stockwell, Edward G. and Bruce Hutchinson. "Mortality Correlates of Economic Status." Population Review, 19 (1975): 45-50.

Stockwell, Edward G. and Karen A. Laidlaw. "Infant Mortality and Socioeconomic Status Among Ohio Counties, 1969-1971." The Ohio Journal of Science, 77 (March, 1977): 72-75.

Struening, Elmer L., et al. "Family, Ethnic and Economic Indicators of Low Birth Weight and Infant Mortality: A Social Area Analysis," Annals of the New York Academy of Science, 218 (June, 1973): 87-107.

Willie, Charles V. "A Research Note on the Changing Association Between Infant Mortality and Socio-economic Status," Social Forces, 37 (March, 1959): 221-227.

Willie, Charles V. and William B. Rothney, "Racial, Ethnic and Income Factors in the Epidemiology of Neonatal Mortality." American Sociological Review, 27 (August, 1962): 522-526.

Woodbury, Robert H. Causal Factors in Infant Mortality. Children's Bureau Publication No. 142. U.S. Government Printing Office, 1925.

Wright, Nicholas H. "Family Planning and Infant Mortality Rate Decline in the United States." American Journal of Epidemiology, 101 (March, 1975): 182-187.

DISCUSSION

Samuel H. Preston, United Nations Population Division

For various reasons I must confine my remarks to the paper by Stockwell, et.al.. This paper provides a useful review and updating of studies of census tract differentials in mortality. Such studies provide one of the cheapest means of documenting the existence of social differences in mortality. The authors are reasonably careful not to push their inferences beyond those which the data can support. They note many of the very serious limitations to which such studies are subject, particularly their inability to provide much detail on the sources of revealed mortality differences. They fail to note one of the advantages of this type of study, namely that its geographic specificity provides a valuable guide to structuring local governmental programmes of health care that are usually implemented on a geographic basis.

There is one very serious disadvantage of such studies, which is partially remediable by improved techniques of analysis. The size of differentials uncovered for one area during one period is not strictly comparable to the size of the differential derived for another population. Stockwell, et.al., attempt to draw inferences about whether differentials are contracting or expanding and whether they are larger in one city than in another. But I doubt that any such inference would be justified without much greater attention to issues of measurement. Take the case of one city in which tract differentials are being compared in 1960 and 1970. If the same tracts form the high and low group in both years, then there is obviously a problem that the social composition of one or both sets of tracts is likely to have changed during the period. If a different set of tracts is used, there is still the problem that the "high" or "low" group may have a quite different mixture of social groups in one year than in another. A tendency toward greater residential intermixture of social groups would obviously tend to produce a contraction of measured differentials between high and low areas, without involving any change in death prospects for individuals. Furthermore, changing the set of tracts can introduce exogenous factors associated with ethnicity, density, access to health care, etc. that will affect measured mortality differentials, without implying any necessary change in underlying relations. In this respect it is wise to remember that census tracts do not have mortality rates, only people do. Tract differentials are valuable only insofar as they are suggestive of individual differences in mortality. As I have suggested, the macro-micro translation problem is acute under present procedures.

It seems to me that a much better way to measure the tract differential would be to use the regression coefficient expressing the relationship between tract death rates and mean tract status on the indicator in question. Such a coefficient would, for example, express the effect of a one-year gain in mean adult educational attainment on the death rate. This effect could then be compared over space and time. No grouping of tracts whatsoever is required. Furthermore, the mixture of groups within tracts would have no effect on the measured differential so long as mortality is linearly related to the characteristic in question. This proposition is easily demonstrated algebraically. Non-linearities will continue to disturb measured relations, as will the occasional need to rely upon medians rather than means. But in general much more confidence could be placed in statements about the relative size of differentials. Such a treatment is readily generalized to one that recognizes various causes of death, since the cause-specific regression coefficients must sum to the regression coefficient for all causes combined. This aggregation property is absent when ratios are employed.

In his earlier review of social differentials in mortality, Antonovsky speculated that after a long period of contraction, differentials may again expand as a result of the development and slow social diffusion of methods of preventing the chronic diseases. Since we seem to have entered at last a period of persistently declining mortality from chronic diseases, it would be interesting to reexamine this proposition. Census tracts are a clumsy vehicle in this regard, but as Stockwell, et.al. quite rightly point out, they are an important stopgap until larger and more expensive studies of individuals are conducted. Census tract studies have served a valuable role in pointing out that social status is still a major dimension of variation in American mortality. When comparably-sized differentials were discovered for a personal habit such as cigarette smoking, there was an enormous outpouring of funds for research to discover the causes and mechanisms of effect. It is unfortunate that there has been no such movement in regard to class differentials. There are some obvious differences related to the specifiability of cause and effect relations. But it is probably also true that the biomedical establishment in the National Institutes of Health is by training and inclination more comfortable supporting studies of physical than of social factors. Demographers using "found" data must continue to call attention to the existence of major social inequalities in the length of life and hope that someone eventually pays attention and supports studies designed to uncover the causes.

STRAIN AT A GNAT AND SWALLOW A CAMEL: OR, THE PROBLEM OF MEASURING SAMPLING AND NON-SAMPLING ERRORS

Tore Dalenius, University of Stockholm

> "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge; but you have scarcely, in your thoughts, advanced to the stage of a science, whatever the matter may be."
>
> -- Lord Kelvin

1. Introduction

The utility of a survey (program) may be expressed in terms of a "utility vector" $U(\cdot)$ with elements:

i. Relevance
ii. Accuracy
iii. Timeliness
iv. Wealth of detail

etc. This paper will focus on the two first-mentioned elements.

It is clear that measures of the relevance and the accuracy are most useful to both producers and users of survey statistics.

Reliable statistics on survey practice are unavailable. It is hoped that the forthcoming ASA-NSF "survey of surveys" will fill this gap in our present knowledge. Lacking such statistics, I will have to base this paper on "trade talk". The opinion appears to be widely held among statisticians and users alike that much of today's survey practice is inadequate for several reasons, three of which are as follows:

a. The relevance is seldom measured. The point seems to be that the relevance is usually taken for granted rather than objectively assessed. It is, for example, not necessarily true that concepts which were adequate 30 years ago, when a given survey program was started, are still adequate.

b. Too often, no (satisfactory) effort is being made to measure the accuracy. Statisticians are typically content to measure the sampling error, while neglecting the non-sampling error. The following quotation, from Wallis (1971), is illuminating:

> "Although there was considerable variation, both for different statistics in the same agency and across agencies, the [Commission's] response to the survey showed disappointingly little knowledge of error structure. Sampling errors were estimated for most statistics based on probability samples, but there were, with only few exceptions, very few analyses of response and other nonsampling errors, even in cases in which, because of long recall or the use of incomplete records, they were likely to be substantial."

c. Measures of the sampling error are too often grossly inadequate. Thus it is not uncommon to use a formula for simple random sampling irrespective of the sampling design actually used. Mention should also be made of the practice of neglecting the fact that what is referred to as a measure of the sampling error may also to some extent reflect response variation.

The survey practice just described may be summarized in terms of "strain at a gnat and swallow a camel"; this characterization applies especially to the practice with respect to the accuracy: the sampling error plays the role of the gnat, sometimes malformed, while the non-sampling error plays the role of the camel, often of unknown size and always of unwieldy shape. There are some signs today that the situation just discussed is worsening: non-response rates have increased significantly in recent years and may become even higher.

If today's unsatisfactory survey practice is not to become tomorrow's malpractice, a radical change is called for. It is the modest purpose of this paper to review the prospects for change and to discuss in general terms an approach to increasing our knowledge about the error structure of surveys, which in my opinion is a sine qua non for that change.

I. BRINGING ABOUT A CHANGE

2. The Notions of "Relevance" and "Accuracy"

The terminology used in discussions of relevance and accuracy is - as shown in Deighton et al. (1977) - characterized by a considerable amount of "linguistic variability", which makes the exchange of ideas and results difficult, to say the least. A necessary though not sufficient condition for bringing about a change is the development of a standard terminology. It is beyond the scope of this paper to suggest standards. I will be satisfied with defining "relevance" and "accuracy" along the lines suggested in Hansen et al. (1964).

The starting point is provided by three basic concepts:

i. the ideal goal $\bar{Z}$
ii. the defined goal $\bar{X}$; and
iii. the outcome of the survey $\bar{y}$.

Using these concepts, it is now possible to define three differences ("errors"):

i. $D(R) = \bar{X} - \bar{Z}$ (reflecting relevance)
ii. $D(A) = \bar{y} - \bar{X}$ (reflecting accuracy)
iii. $TD = \bar{y} - \bar{Z} = D(R) + D(A)$ (reflecting total difference)

3. A Plan for Bringing About a Change

The change I envision is to make it a rule, not an exception, that relevance and accuracy are measured.

It is clearly no easy matter to bring about such a change; it may call for many years of hard work. While there may be several alternative courses which would accomplish the same end, I will expand upon one specific one here.

The plan takes as its starting point (my conception of) the mechanisms (sources of errors) which generate the differences D(R) and D(A). Some resulting contributions to these differences are of a random nature and may thus be modeled by means of random variables and measured in terms of variances. Other contributions are of a systematic nature; they must be measured in terms of biases. The plan calls for reducing the biases even at the possible expense of increased variances. This idea is, of course, not new; it has long been used, for example, at the U.S. Bureau of the Census (Hansen et al. (1967)). The rationale of the plan is that it is usually much easier to cope with random errors than with systematic errors.

In sections 4 and 5, I will discuss how this plan may be applied to the control and measurement of the relevance and the accuracy, respectively.

4. Control and Measurement of the Relevance

Whether the statistics to be produced are classified as "general-purpose" or "special-purpose", the design of a survey must clearly reflect some specific purposes. The statistician must take into account who the potential users are and what the problems are to the solution of which they expect the survey to contribute.

The design procedure can indeed be formalized in a way that should enhance the control and measurement of D(R). I will dwell upon one such formalization here.

4.1 Control of D(R)

Consider a group of potential users with related or similar problems. Associated with this group, there is a set of ideal goals:

$$\bar{Z}_1, \bar{Z}_2, \ldots, \bar{Z}_j, \ldots, \bar{Z}_k$$

Corresponding to these ideal goals, there is a set of feasible defined goals:

$$\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_i, \ldots, \bar{X}_h$$

where typically $h \leq k$.

For each pair $\bar{X}_i$, $\bar{Z}_j$, there is a difference:

$$D_{ij} = \bar{X}_i - \bar{Z}_j$$

which reflects the relevance of $\bar{X}_i$ vis-à-vis $\bar{Z}_j$. This difference may be exhibited as a matrix:

$$[D_{ij}] = \begin{bmatrix} D_{11} & \cdots & D_{1j} & \cdots & D_{1k} \\ \vdots & & \vdots & & \vdots \\ D_{i1} & \cdots & D_{ij} & \cdots & D_{ik} \\ \vdots & & \vdots & & \vdots \\ D_{h1} & \cdots & D_{hj} & \cdots & D_{hk} \end{bmatrix}$$

If this matrix were known prior to the survey, it could be used to select the defined goal which in some sense is best. But by the same token, there would then be no need for the survey!

What is needed, obviously, is some method for approximating the matrix.

4.2 Approximating $[D_{ij}]$

I will point to two possible ways of approximating $[D_{ij}]$. One way calls for replacing the elements D_{ij} by "preference scores" P_{ij} reflecting the preferences among the users as to the pairs $\bar{X}_i$, $\bar{Z}_j$. Another way calls for replacing the elements D_{ij} by indicators a_{ij}, where $a_{ij} = 1$ if the pair $\bar{X}_i$, $\bar{Z}_j$ is judged by the users to be acceptable, and otherwise $a_{ij} = 0$.

A matrix $[P_{ij}]$ or $[a_{ij}]$ can be analyzed and assessed as a basis for selecting the defined goal (which is not necessarily one of the originally conceived defined goals). The analysis and assessment may be carried out along the lines discussed in Dalenius (1968). A condition for this procedure to be feasible and useful is obviously that intimate cooperation be established between the statistician and the users. The statistician must take an active role in getting to understand the users' problems; by the same token, the users must learn to understand the ramifications of alternative choices of defined goal. In addition, the computing expert must take an active part in the design.

In some cases - perhaps more often than not - it may not be feasible to approximate $[D_{ij}]$ as suggested above. In these cases, the construction of an error profile, to be discussed in part II, may provide some helpful insights.

5. Control and Measurement of the Accuracy

The difference D(A) may be written:

D(A) = sampling error + non-sampling error

where the sampling error is relative to the outcome of "equal complete coverage" (Deming (1960)) and the non-sampling error accounts for the balance of D(A).

5.1 Control of D(A)

While control should be aimed at both components of D(A), it seems especially important to focus on the non-sampling component. The error profiles presented in Bailar and Brooks (1977) and Madow (1977) support that contention. In what follows, I will discuss two possible approaches to control of the non-sampling component.

a. The first approach calls for identifying survey operations with high risks for deviation between design and execution which are difficult to control in a satisfactory way; (some of) these operations may then be replaced by operations with low risks. In some cases, this will mean replacing a "complicated" operation by a "simple" one, especially if the complicated operation is primarily a human operation (like coding). In other cases, the action to take will be the reverse one: a "simple" human operation is replaced by a "complicated" automatic (computerized) operation. Editing is an example of an area in which this idea is already successfully applied.

b. The second approach calls for (better) monitoring of the survey operations. This may necessitate the development of a special signal system which helps to identify problems while there is still time to take "preventive action". Non-response is an example of a kind of problem in which this approach should be relatively easy to apply in all surveys.

5.2 Measurement of D(A)

It is worth noting that theory and methods are in fact available for this measurement.

a. In the context of what has become known as the U.S. Bureau of the Census survey model, theory is developed for measuring D(A) by the mean-square error of the estimator:

MSE = sampling variance + response variance + interaction + squared bias

as discussed in Hansen et al. (1964). Moreover, methods are available - see Bailar and Dalenius (1969) for a systematic account - for the design of schemes which may be used to estiamte the components of the MSE.

b. In Lessler (1974), theory and methods are available for using two-phase sampling as a means of controlling and measuring D(A).

Thus, the fact that D(A) is seldom (adequately) measured cannot be explained by lack of the tools necessary for doing so. While there may be many reasons for the current state of affairs, I presume that cost considerations often play a decisive role.

At any rate, it is likely that a change will not take place spontaneously; it will have to be generated. One way of stimulating the change may be to illuminate the importance of measuring D(A) by means of some "second-best" measurements:

i. Measuring the "representativity" by comparing survey estimates with known population characteristics. This type of measurement dates back to the early days of purposive selection, and was often used in an uncritical way.
ii. Applying "error ratio analysis" as suggested by Brown (1967).
iii. Computing "quality codes" as developed by Zarkovich (1967).
iv. Constructing an "error profile".

In part II, I will dwell on this fourth option; I will in fact argue that it may serve a useful purpose with respect to both D(R) and D(A).

II. THE ERROR PROFILE APPROACH

6. The Notion of an Error Profile

Hansen et al. (1967) discuss what to do in a situation in which it is not feasible to measure D(A) by means of the mean-square error. In essence, they suggested that the statistician provide for a disclosure of the survey operations.

The term "error profile" will be used here in a way consistent with that suggestion; this term is chosen in preference to the longer though somewhat more appropriate term "profile of sources of errors". More specifically, an error profile is a systematic and comprehensive account of the survey operations which yield the statistic $\bar{y}$ and thus the differences D(R) and D(A).

Constructing an error profile calls for assessing each survey operation with respect to:

i. The presence or absence of a deviation between design and execution.
ii. The size of this deviation.
iii. The impact of this deviation; as a special case, this impact may be expressed in terms of a contribution to the MSE.

It should be remarked that it may not be possible to assess each survey operation with respect to all three elements just listed.

There is no standard format for an error profile. I will mention here two possible formats:

i. One format is based on a list of the survey operations in the order in which they were executed.
ii. The second format calls for assigning the survey operations to homogeneous groups on the basis of the purpose of each operation.

The second format has, it seems to me, the advantage of lending itself to some standardization. In section 7, I will present one possible grouping scheme.

7. A Possible Grouping of the Survey Operations

The starting point is the total difference:

$\bar{y} - \bar{Z} = D(R) + D(A)$

Against the background of this difference, I identify a hierarchy of survey operations:

i. Primary Survey Operations - PSOs
ii. Secondary Survey Operations - SSOs
iii. Tertiary Survey Operations - TSOs, etc.

In the interests of being specific, I will give a couple of illustrations.

It seems natural to distinguish two PSOs:

PSO-1: Design of the survey
PSO-2: Execution of the design

PSO-1 may be divided into SSOs as follows:

SSO-11: Choice of the properties to measure
SSO-12: Choice of the survey population

while PSO-2 may be divided into SSOs as follows:

SSO-21: Getting observational access to the population: developing the frame, selecting the sample, etc.
SSO-22: Collecting the data
SSO-23: Processing the data
SSO-24: Computing the survey statistic $\bar{y}$
SSO-25: Computing measures of D(R) and D(A)

8. The Assessment Procedure

As mentioned in section 6, constructing an error profile calls for assessing each survey operation with respect to three aspects: presence/absence of a deviation between design and execution; size; and impact. The procedure to use for this assessment will, of course, depend upon the nature of the survey operation to be assessed. I will limit myself here to giving two minor illustrations.

8.1 Assessing PSO-1: Design of the Survey

As discussed in section 7, PSO-1 may be divided into two SSOs:

SSO-11: Choice of the properties to measure
SSO-12: Choice of the survey population

I will discuss the assessment of these SSOs in turn.

a. Corresponding to the defined goal $\bar{X}$, there is a survey variable X defined by reference to:

i. the property to measure
ii. the measurement method

Similarly, corresponding to the ideal goal $\bar{Z}$, there is an ideal variable Z defined in the same way.

The analysis of the choice of the survey variable calls for determining whether the survey variable is equal to the ideal variable with respect to "property to measure" and "measurement method": X = Z, or whether it differs from the ideal variable: $X \neq Z$.

In a specific survey, X and Z may be defined by reference to the same property, but they may differ with respect to the measurement method. As an example, the measurement method corresponding to Z may not be operationally feasible for the survey under consideration.

If the analysis shows that $X \neq Z$, there is a "definitional bias" associated with the defined goal $\bar{X}$.

b. Corresponding to the defined goal $\bar{X}$, there is a population of objects - the survey population - to be denoted by $[O(\bar{X})]$; technically, it is represented by the frame.

Similarly, corresponding to the ideal goal $\bar{Z}$, there is a population of objects - the target population - to be denoted by $[O(\bar{Z})]$.

If all objects in $[O(\bar{X})]$ are also in $[O(\bar{Z})]$, and all objects in $[O(Z)]$ are also in $[O(X)]$, the survey population is equal to the target population:

$$[O(\bar{X})] = [O(\bar{Z})]$$

If some objects in $[O(\bar{X})]$ are not in $[O(\bar{Z})]$, or some objects in $[O(\bar{Z})]$ are not in $[O(\bar{X})]$, the survey population is different from the target population:

$$[O(\bar{X})] \neq [O(\bar{Z})]$$

In fact, this latter situation is the typical one in applications. It calls for assessing the difference between $[O(\bar{X})]$ and $[O(\bar{Z})]$ by comparing these populations with respect to:

i. The rules associating objects with $[O(\bar{X})]$ and $[O(\bar{Z})]$, respectively;
ii. The (approximate) frequencies:

N_{11} = the number of objects which belong to both $[O(\bar{X})]$ and $[O(\bar{Z})]$
N_{10} = the number of objects which belong to $[O(\bar{X})]$ but not to $[O(\bar{Z})]$
N_{01} = the number of objects which belong to $[O(\bar{Z})]$ but not to $[O(\bar{X})]$

The ratio:

$$R = \frac{N_{11}}{N_{11} + N_{01}}$$

may be looked upon as a measure of the appropriateness of $[O(\bar{X})]$.

The point just made about R may be illustrated by considering the case of a survey which yields an estimate $t = pN_{11}$ of the target population total:

$$T = P(N_{11} + N_{01})$$

where P is, for example, the rate of unemployed persons. If R is close to 1, then t is close to T (granted that p is close to P).

8.2. Assessing TSO-221: Observing the Objects Selected for the Survey

In real-life surveys, the number of TSOs is likely to be large. I will select one of them - TSO-221 - for illustration: observing the objects selected for the survey (irrespective of the method of operation).

In practice, it will happen that some objects become "non-respondents". It is in principle relatively simple to measure the size of this specific event; this does not mean, however, that it is adequately done in all instances. As to measuring the impact of the non-response, it is in some cases (notably when $\bar{X}$ is a proportion) possible to compute an upper and lower value for this impact.

9. The Error Profile Documentation

A comprehensive account of the assessment of survey operations may possibly become a rather sizeable document, especially if it is to be self-contained and deals with a survey which is not repeated. It may therefore prove desirable to try to summarize these findings in a simple error profile protocol, or table, the headings of which may be as in figure 1 below.

Survey operation	Kind of deviation	Size	Impact
PSO-1: Design of the survey			
SSO-11: Choice of properties to measure			
SSO-12: Choice of survey population			
PSO-2: Execution of design			
PSO-21: ...			
PSO-22: ...			
etc.			

Figure 1.

10. Limitations and Potentialities of the Error Profile Approach

In sections 4 and 5, constructing an error profile was suggested as a means of measuring D(R) and D(A).

In section 6, I defined an error profile to be "a systematic and comprehensive account of the survey operations which yielded the statistic $\bar{y}$ and thus the differences D(R) and D(A)."

The error profile approach is as yet virtually untested. Thus, it would be premature to pass

any judgment on its usefulness; the proof of the pudding is in the eating.

The main limitation of the error profile approach is obvious: it does not make it possible to measure the components of the mean-square error of $\bar{y}$ within the framework of some survey model. On the other hand, the limited experiences as yet available support the contention that it has some significant potentialities. Thus, the error profile approach:

i. encourages comprehensive documentation of the survey operations;
ii. helps to identify "error-prone" survey operations; and
iii. serves as a summary protocol of research and development already carried out and yet to be carried out.

Acknowledgement: This paper reflects in several ways discussions within the Subcommittee on Non-Sampling Errors, organized by the Federal Committee on Statistical Methodology, Statistical Policy Division (Office of Management and Budget).

References

Bailar, B.A. and Dalenius, T. (1969): Estimating the response variance components of the U.S. Bureau of the Census' survey model. Sankhya, B, 341-360.

Brooks, C.A. and Bailar, B.A. (1977): An error profile: employment as measured by the current population survey. Paper presented at the 137th Annual Meeting of the American Statistical Association, Chicago, Ill., August 15-18, 1977.

Brown, R.V. (1967): Evaluation of the total survey error by error ration analysis. Metra, 593-613.

Dalenius, T. (1968): A feasible approach to general purpose sampling. Management Science, 110-113.

Deighton, R.E., Poland, J.R., Stubbs, J.R. and Tortora, R.D. (1977): Glossary of nonsampling error terms. Paper presented at the 137th Annual Meeting of the American Statistical Association, Chicago, Ill., August 15-18, 1977.

Deming, W.E. (1960): Sample design in business research. Ch. 4. John Wiley & Sons, Inc., New York.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1964): The estimation and interpretation of gross differences and the simple response variance. In Rao, C.R. (editor): Contributions to statistics presented to Professor P.C. Mahalanobis on the occasion of his 70th birthday. Pergamon Press, Oxford, and Statistical Publishing Society, Calcutta.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1967): Standardization of procedures for the evaluation of data: measurement errors and statistical standards in the Bureau of the Census. Paper presented at the 36th Session of the International Statistical Institute in Sydney, 1967.

Lessler, J.T. (1974): A double sampling scheme model for eliminating measurement process bias and estimating measurement errors in surveys. Institute of Statistics Mimeo Series No. 949, University of North Carolina, Chapel Hill.

Madow, L.H. (1977): An error profile: employment as measured by the current employment statistics program. Paper presented at the 137th Annual Meeting of the American Statistical Association, Chicago, Ill., August 15-18, 1977.

Wallis, A. (Chairman) (1971): The President's Commission on Federal Statistics. Vol. II. Government Printing Office, Washington, D.C.

Zarkovich, S.S. (1967): A system of statistical quality codes. Paper presented at the 36th Session of the International Statistical Institute in Sydney, 1967.

AN ERROR PROFILE: EMPLOYMENT AS MEASURED BY THE CURRENT POPULATION SURVEY

Camilla A. Brooks and Barbara A. Bailar
Bureau of the Census

I. Introduction

A. Purpose of the Error Profile

Ideally, a user of survey data should be provided two measures of error with each statistic produced from a survey--one, a measure of the total variance, including sampling, response, and processing variability and second, a measure of the total bias, including the bias arising from the data collection procedure, the questionnaire used, and the estimator. These two measures are rarely, if ever, available. An estimate of the sampling variance, which may or may not adequately reflect all sources of variance, is all that is usually given to a user to evaluate survey data.

The interests of survey designers and some data users, however, extend beyond this to the individual components of variance and bias and how they may interact. One of the most studied surveys sponsored by the Federal Government is the Current Population Survey (CPS). Studies have been conducted to assess the impact of many of the aspects of the survey design on the estimates. Yet, even with this survey, the specific effect of each of the various sources of bias and variance on the survey statistics and their interaction are not fully quantified. This paper gives an illustration of how one can go about constructing an error profile for a survey statistic by examining each of the potential sources of nonsampling error and trying to assess the impact of each source on the survey data. These assessments would be input to a mathematical model for the total survey error. To illustrate the impact of nonsampling error, we have drawn upon a number of studies by Census Bureau staff members and others. We did not attempt to construct a mathematical model for the total error in a survey statistic. However, one of the main conclusions that can be drawn from this illustrative error profile is that much more empirical work is necessary to provide a comprehensive picture of the effects of nonsampling error on a survey statistic.

The focus of this paper is on the employment statistic. Ideally, one would list each potential source of nonsampling error and then quantify the effect of that source on the employment statistic. Because data as well as space are limited, the major steps in the survey process are mentioned, but only a few steps are described in detail. For a more thorough explanation of the Current Population Survey in its entirety the reader should refer to Hanson's draft revision of Technical Paper No. 7 (1976) which should soon be published.

B. Specifications of the CPS Employment Statistic

The concept of "employment" used in the CPS is specified by the Bureau of Labor Statistics (BLS), and its translation into a set of questions for collecting the data is carried out jointly by BLS and the Census Bureau. Since January 1967 the definition of employment used in the CPS has been: Those at work, consisting of persons who worked one hour or more for pay or profit, or 15 hours or more without pay in a family operated enterprise, and those with a job, but not at work, such as persons temporarily absent from work because of illness, vacations, etc.

The CPS is restricted to the civilian noninstitutional population age 14 and over. A monthly housing unit sample is used which represents the universe of all households in the United States and also includes nonhousehold units, in which people live such as hotels, dormitories, flophouses, bunkhouses, and the like. In using the housing unit approach, an implicit assumption is made that each person 14 years of age and over is uniquely associated by the survey definitions and procedures with either a household or one of the nonhousehold units mentioned.

C. The Survey Design

The CPS is redesigned after each decennial census in order to utilize the most recent census data. In the 1970 redesign, which is discussed here, 461 primary sampling units (PSU's), which are primarily counties or groups of counties, were selected from 376 strata. Of these 461 PSU's 156 were designated self-representing (SR). The other 305 PSU's were selected from 220 nonself-representing (NSR) strata and are referred to as nonself-representing PSU's.

The CPS is a multi-stage cluster sample which is essentially self-weighting. The design is such that the ultimate sampling units (USU's)-that is, clusters of approximately four, usually contiguous, housing units, do not remain in sample for the entire decade. Therefore, several CPS samples must be generated for use during the decade. Each CPS sample consists of eight, approximately equal, systematically selected subsamples known as rotation groups. These rotation groups are introduced into the sample once a month for eight months using a 4-8-4 rotation scheme; i.e., each sample USU is in sample four months, out eight months, and then in four more months. Under this scheme, each month there is a 75 percent month-to-month overlap of households and a 50 percent year-to-year overlap. At the time of the 1970 redesign, data were collected from approximately 47,000 eligible households each month.

II. The Sampling Frame

One potential source of bias in any sample survey is the lack of complete coverage of the population. The coverage is associated with the frame, and, if the frame is deficient, or information is not obtained from all persons within the sample units, there will be undercoverage.

A. Description of the Frame

The frame for this survey is derived from a variety of sources with the main source the 1970 Decennial Census. In the CPS extensive use is made of the 229,000 enumeration districts (ED's) defined in advance of the census; these are large geographic areas, each containing about 350 housing units on the average. There were three types of ED's in the census identified by the manner of forming the address register which was a list of the housing units within an ED. These are as follows:

1. Tape address register (TAR) ED's in which the address register was created from a computer tape copy of a commercial mailing list and corrected by the Post Office and local agencies.
2. Prelist ED's in which the address register was constructed by a listing procedure conducted in advance of the census.
3. Conventional ED's in which the address register was prepared by the enumerators during the enumeration.

For purposes of the CPS sample, a 1970 Census ED is referred to as an address/list ED if the conditions listed below are satisfied.

1. The ED is a TAR ED.
2. The ED is a prelist or conventional ED satisfying "a" and "b" below:
   a. at least 90 percent of the 1970 Census addresses within the ED are recorded with complete street name and house number;
   b. the ED is located in an area which issues building permits.

In address ED's the CPS sample is selected from the census address registers and the resulting sample referred to as an address sample. The address ED's are supplemented by the Census supplemental sample, referred to as the Cen-Sup sample. which is used to cover housing units in address ED's at addresses missed in the census or inadequately described in the address register. These units represent less than one percent of the CPS sample. All other 1970 Census ED's are referred to as area ED's. These are subdivided into area segments which are sampled and assigned to be listed by an enumerator about a month before interview. The sampling of housing units from the listings is carried out in an office operation, not by the interviewer.

Units built after April 1, 1970 are represented in address ED's and permit issuing area ED's by a sample of building permits selected from records of places issuing building permits as of January 1, 1970. This supplement to the frame is referred to as the permit universe and includes approximately 10 percent of the CPS sample. In non-permit issuing area ED's new construction is covered by interviewer listing.

B. Potential Sources of Errors Associated with the Sampling Frame

It is known that the sampling frame used for the CPS does not fully represent the universe of housing units. However, deficiencies discussed below are estimated to represent less than three percent of the target population of persons. Some of the frame deficiencies are discussed below.

Permit Lag Universe

In permit issuing ED's, housing units completed after the census for which permits were issued before January 1, 1970 are not included in the CPS frame. These units are collectively referred to as the permit lag universe. There is an estimated total of 598,000 units for which permits were issued prior to January 1, 1970 that were completed after April 1, 1970 (MacKenzie, 1977).

Undercoverage of Special Places - Mobile Homes

Mobile homes located in address segments are a second potential source of coverage loss. Presently, in the CPS there is no general procedure for identifying or representing mobile homes in new mobile home parks, or new mobile homes at large in address ED's at addresses nonexistent in the 1970 Census; the permit universe includes regular housing units only. In addition to new mobile homes, the coverage problem of mobile homes in address ED's extends to those in existence at the time of the census but not counted in the census, and to those vacant in 1970 and therefore by design not counted in the 1970 Census but which are now occupied.

The coverage improvement program in the October 1976 AHS located approximately 300,000 mobile homes for the period April 1970-October 1976. This improvement program has not yet been included in the CPS, but these mobile homes should be represented in the sample. Though concern is greater for mobile homes, other special places including transient hotels, boarding homes, etc. could present some of the same problems as the mobile homes.

Nonpermit Issuing TAR ED's

A small number of the TAR ED's (approximately 47-50 or about 0.3 percent of all TAR ED's) are in non-permit issuing areas. Because of irreconcilable problems in sampling them as area ED's, it was decided to treat these ED's in the same manner as permit issuing, address type ED's. Thus new construction in these ED's is not represented (Baer, 1973 and Boisen, 1971).

Other Structure Misses

Other problems with the frame in address ED's not addressed by the Census Supplemental sample include coverage of homes moved to a site with an address not in the 1970 Census and structures converted from nonresidential to residential use after the census.

C. Errors Associated With the Coverage of Persons Within Sample Housing Units

Within household coverage misses and the undercoverage of individuals with no attachment to any address are believed to account for a large percentage of the frame deficiencies;

however, information on both the extent and the causes of this problem is limited. It is estimated that because of missed structures less than three percent of the target population is not included in the frame. However, Table 4 in Section V shows ratios of independent estimates of the population prepared by the Census Bureau to Current Population estimates of the population which indicate a coverage problem exceeding three percent. For white males and females the ratios are 1.049 and 1.023, respectively while for males and females of black and other races the respective ratios are 1.155 and 1.075. Further, by agreement with the data users the independent estimates of the population which are used in Table 4 do not reflect the estimated undercoverage of the census which itself did not include an estimate for the "illegal alien" population in the U.S. (Siegal, 1973). Thus, the within household coverage problem is even greater than indicated in the tables. Of blacks missed in the 1970 Census an estimated 64 percent were missed within units enumerated as occupied or occupied units enumerated as vacant; the corresponding figure for persons of white and other races was 42 percent (Jones and Blass, 1975).

This phenomenon of undercoverage compared to the decennial census, which itself is subject to undercoverage of the population, is typical of household surveys generally.

D. The Effect of the Census Undercount on the Employment Statistics

It was noted that, although estimates of the undercoverage of the population by age-sex-race in the 1970 Census are available, the CPS coverage is adjusted to the level of the 1970 Census rather than to figures adjusted for the 1970 undercoverage. A study by Johnston and Wetzel (1969) explored the effect of the 1960 Census undercount on the average monthly labor force estimates for 1967 if this convention were not adopted. Since the labor force status of the omitted persons is unknown, the authors provided two alternative sets of "corrected" labor force estimates. In the first set, an assumption was made that the missed persons had the same labor force status as their peers (persons of the same age, sex, and race group). This is called the "comparability" assumption. In the second set, omitted persons were assumed to have labor force status comparable to people of the same age, sex, and race but living in urban poverty areas. This is called the "poverty neighborhood" assumption.

Their report shows that the most significant change was in the level of employment which would increase by 2.8 million under the comparability assumption and 2.7 million under the poverty neighborhood assumption. The report further shows that the effect of the undercount on labor force estimates is much more important for persons of black and other races than for whites because their undercoverage is greater.

The conclusion of the authors was: the population undercount leads to a substantial understatement of employment levels, and introduces further complications due to a discrepancy in level when these data are related to other time series not based on household surveys. The authors make no comment on the impact of this phenomenon for comparison of changes in level on either a current (short term) or historical (long term) basis.

Hirschberg, Yuskavage, and Scheuren (1977) have recently arrived at estimates of the effect of CPS and 1970 Census Undercoverage on the labor force estimates which are more sensitive to the coverage problem than those of Wetzel and Johnson. These are presented in a paper prepared for this year's ASA meetings.

III. The Data Collection Procedure

A. Description of the Process

The data collection procedure must be both quick and accurate. Interviews are conducted during the week (Monday through Saturday) containing the 19th day of the month, designated interview week. During interview week interviewers obtain information from respondents regarding the previous week or week containing the 12th day of the month, designated survey week.

The data collection procedure includes several potential sources of nonsampling error. For example, perhaps the way the question is worded is not clear to respondents; the use of proxy respondents for household members not at home at the time of the interview may affect the data; the use of telephone interviewing for certain households may change the kinds of data collected; the training of the interviewers may affect the way they collect or record the data. In this section of the paper, attention is focused on the actual conduct of the interview as a potential source of error.

Though any responsible adult household member 14 years of age or older is eligible to act as respondent, the interviewer is encouraged to interview the most knowledgeable household member, usually the household head or wife of head.

At the initial visit the interviewer records on a control card the name, usual residence, relationship to head, date of birth, age, race, sex, etc. for each person in the household. At each subsequent visit to the household the listing is updated. The CPS questionnaire is then completed for each household member 14 years old or older.

B. Potential Sources of Error Associated with the Conduct of the Interview

Mode of Interview

Before the use of the telephone was instituted in CPS interviewing, a test in a limited number of PSU's was conducted to determine its effect on the data. This test, conducted in the early 1950's, showed no appreciable difference in the labor force data obtained by the two methods of interviewing, personal visit and telephone (Hanson, 1976). However, the test conducted at the time was not a completely controlled experiment and the results for today's purposes are outdated. Not only has telephone interviewing increased, but attitudes have probably changed over the years. Because of

the wide use of telephone interviewing in the CPS, there is growing concern about its possible effects on the data, and because of this, studies are now being planned to learn more of its effects.

Interviewers are allowed to interview by telephone according to the following regulations:

> Interviewers are instructed to conduct the first and fifth month interviews by personal visit. Second month interviews may be conducted by telephone only if no one was home at the initial personal visit. Providing the household has consented, interviews in the other months may be conducted by telephone.

The figures discussed here represent 1976 averages of telephone interviews. Though interviewers are instructed not to conduct first and fifth month interviews by telephone, 2.8 and 10.3 percent, respectively, of first and fifth month interviews were, in fact, collected by telephone, presumably to reduce the nonresponse rate that would have been experienced otherwise. Also, 44.5 percent of second month interviews and 76.0 percent of the interviews conducted in the remaining months were collected by telephone.

At the present time there is no evidence that personal interviewing and telephone interviewing yield different results on employment questions. However, it is recognized that the use of the telephone may cause a different respondent to be interviewed. Data show that there is an increase in the number of "other relatives" who are respondents in later months in sample. To the extent that "other relatives" in the household may not be as knowledgeable as the head and/or wife about the labor force status of all household members, the telephone data may not be as accurate.

The Use of Proxy Respondents

In the CPS proxy respondents are frequently used; e.g., the wife may frequently respond for her husband, or in the case when both the head and his wife are absent, a 14 year old may respond for the entire household.

Between February 1965 and June 1966 a Methods Test was conducted outside of the regular CPS but with the purpose of testing new methods for the CPS. One of the problems evaluated in the test was the selection of best respondent for individual household members. Two different studies were conducted. In the first, three procedures were compared as follows:

a. Procedure 1 was similar to the present CPS procedure in that any responsible household member was accepted as a respondent for the entire household;
b. In Procedure 2 each adult household member was to be interviewed for himself;
c. In Procedure 3 an advance form containing important labor force questions was sent to each household in the test with a request that each adult household member fill the form for himself. The interviewer was then to transcribe this information to the questionnaire and ask the household respondent the remaining questions about the household members.

A comparison of the results was provided in a memorandum by Deighton (1967). The percentage of persons employed as measured by procedures 1, 2, and 3 were 55.6, 57.2, and 57.3, respectively. The self-respondent procedure resulted in more persons employed than did the household respondent procedure by approximately 1.6 percentage points. In a second experiment reported by Williams in 1969 similar results were found. However, sampling errors were not provided in either case so it is questionable whether this difference was significant.

Influence of Interviewers

Interviewers have many opportunities to influence the data. In this paper the focus will be on the noninterviews and misclassification of the labor force status.

The interviewer may encounter three types of noninterview situations: Type A noninterviews - those households eligible for the survey for which the interviewer was unable to complete the interview; Type B noninterviews - vacant units, vacant sites, or units occupied by persons ineligible for the survey; and Type C noninterviews - units demolished, converted to permanent storage or business use, moved from site, or found to be in sample by mistake. Interviewers are strongly urged to keep the Type A noninterviews to a minimum. Type B noninterview households are visited each month to determine if any have become eligible for interview. Type C noninterview units are not visited again by the CPS interviewer.

Table 1 shows the noninterview misclassifications for the years 1973-76 as determined from reinterviews of subsamples of housing units in the CPS. There was an annual misclassification rate of 2.4 percent for 1976 which was significantly different from the 1974 rate only. The table shows that there are more Type A noninterviews that are misclassified as B's or C's than the reverse. This would lead to a deficit in the number of households eligible for interview, and therefore, in the coverage of the target populations and may explain part of the observed undercoverage referred to earlier.

Another aspect of the impact of interviewers is in the labor force status assigned to individuals on the basis of their responses to the survey questions. Accuracy of the data collected by the interviewer is frequently measured by using the results of the monthly CPS reinterview survey as a standard. The CPS reinterview is conducted by senior interviewers or members of the supervisory staff, and as such, are considered to be more accurate than the original interview results; however, because of limitations of the reinterview survey its results should by used with caution.

Specific to the number of persons classified as employed in the CPS, the reinterview provides information on how many persons were classified as employed in the reinterview. If one is willing to accept the reinterview as a standard, then the difference between the original interview and the reinterview can be used as a measure of bias. Table 2 shows the results of the two estimates of employment annually

Table 1. Noninterview Misclassification Rates[1/]
(Percentages)

Misclassification	1973[3/]	1974[3/]	1975[3/]	1976[3/]
Total	2.7	3.3	2.9	2.4
A's as B's	1.4	2.1	1.8	1.6
B's as A's	0.7	0.4	0.2	0.4
C's as B's	0.45	0.7	0.6	0.4
Other 2/	0.15	0.1	0.3	0.03

1/ Moye, 1976 and Schreiner, 1977
2/ B's as C's, C's as A's, and A's as C's
3/ Base - total noninterviews

from 1956 through 1976. From 1956 to 1960 the reinterview was 0.2 percentage points lower than the original interview. However, since that time the differences have increased with a high in 1976 of 0.7 percentage points. Though these differences may seem small, they are consistent and when applied to an employment figure of 80 million, they account for between 160,000 to 560,000 persons.

Table 2. Summary of Percent of Persons Employed as Measured in the Original CPS Interview for the Reinterview Subsample and as Measured by the Reinterview after Reconciliation, 1956,1976.

Year	Percent of Persons in labor force employed: Original	Reinterview	Reinterview minus original
1956	96.3	96.1	-0.2
1957	95.8	95.8	0.0
1958	93.2	93.0	-0.2
1959	94.4	94.2	-0.2
1960	94.6	94.4	-0.2
1961	93.1	92.8	-0.3
1962	94.5	94.5	0.0
1963	94.4	94.0	-0.4
1964	94.8	94.3	-0.5
1965	94.9	94.7	-0.2
1966	96.1	95.8	-0.3
1967	96.2	95.8	-0.4
1968	96.3	96.0	-0.3
1969	96.3	95.9	-0.4
1970	94.9	94.5	-0.4
1971	94.1	93.7	-0.4
1972	94.7	94.4	-0.3
1973	95.0	94.7	-0.3
1974	94.5	93.9	-0.6
1975	91.8	91.2	-0.6
1976	92.5	91.8	-0.7

IV. Data Processing Operations

There are many activities included in the processing of the CPS questionnaires. These operations include the coding of occupation and industry, the microfilming of the questionnaires, the conversion of the microfilm to computer tape by means of a process called FOSDIC (Film Optical Sensing Device for Input to Computers), the machine editing of the data and imputation for missing values. Though all of these activities are potential sources of error, in this paper we discuss only the FOSDIC operation.

A. Description of the FOSDIC Operation

"FOSDIC....can be described as a machine which is capable of 'reading' information from a microfilm copy of an appropriately designed schedule and transferring this intelligence to magnetic tape for processing on electronic computers" (McPherson and Volk, 1962).

Several variables play a role in the microfilming and FOSDIC procedures and their adherence to standards can determine the success or failure of the data processing. These variables include, but are not limited to, the quality of the paper used for the questionnaires; the uniformity of the index marks and marking circles on the questionnaire; and, of course, the proper operation of the FOSDIC and microfilming equipment.

B. Potential Sources of Error in the FOSDIC Operations

Distributions of CPS questionnaires rejected by FOSDIC by cause of error are regularly prepared. In January 1976 out of 72,172 total questionnaire forms 1,074 or 1.49 percent of the questionnaires were rejected (Jablin, 1977). Of these 1,074 questionnaires, 379 or 0.5 percent were rejected because of FOSDIC/filming errors. The errors in FOSDIC/filming that are associated with filming problems, bad index marks, etc. are corrected; for example, if the questionnaire is "bad", data are transcribed to another questionnaire, questionnaires with missed pages are remicrofilmed, etc. The errors that represent a threat to data accuracy are the "invisible errors", i.e. errors that cannot be detected. An example is a FOSDIC pickup of an incomplete erasure as a mark. Also, FOSDIC itself is subject to a certain amount of measurement error; it is possible with the same tolerance levels for reading marks, that it can get different readings for the same marks read at different times.

CPS/FOSDIC Study

In 1974-75 a CPS/FOSDIC Study was conducted in the Operations Analysis and Quality Control Branch of the Bureau to identify specific sources of variations in the system. One major aspect of the CPS/FOSDIC Study involved the reading of two identical pages (both containing labor force data) of 300 CPS questionnaires. The two pages were filled in identically for each of the 300 documents so that there were then 600 identical pages of information. The use of different cameras for filming and different FOSDIC readers produced 32,997 attempted readings.

Out of the 32,997 attempted readings, the following errors were observed (Boisen, 1975):

Table 3. Some Results of the FOSDIC Experiment

Problem Area	Number	Percent of Total
Drops of marks	22	.0029
Pickups of blanks	44	.0034
Drop-and-pickups	1	.00013
Skipped read areas	260	.013
Skipped pages	27	.082

Some of the major findings resulting from this aspect of the CPS/FOSDIC experiment were that (1) basically the system as operated during the experiment was under control with system error so slight that improvement could be impossible; (2) quality control procedures should be extended to the marking of CPS questionnaires; and that (3) further investigation might pinpoint some nonrandom and significant sources of error that result in failed calibrates and missed indices.

FOSDIC error is quite small and represents a gain in accuracy when compared with the previous use of keypunching of the data. Thus processing should have less effect on the accuracy of surveys using this procedure.

V. Preparation of Estimates

There are several stages in the preparation of the final estimates of employment. There is a weighting procedure which attempts to adjust for noninterview and undercoverage, a series of estimates culminating in what is known as a composite estimate, the seasonal adjustment of the point estimates, the estimates of sampling variance, and the estimates of nonsampling error. In this paper only the weighting procedure is discussed.

A. The Weighting Procedure

The sample as selected for the CPS is essentially self-weighting, i.e. each sample household has approximately the same probability of selection. However, because of nonresponse and coverage problems, a reweighting of the records occurs before the estimates are produced. In the CPS there are five distinct steps in the reweighting process. These are as follows:

1. The reciprocal of the probability of selection is attached to the record for a given unit.
2. During listing it was discovered that certain USU's contained far more units than were expected based on census listing so subsampling took place to make the interviewer workload manageable. The weights for the subsampled units are now multiplied by a factor which inflates these units to reflect the actual number of units in the USU. This process is called duplication control. The maximum factor used in the duplication control procedure is four. When USU's are unusually large (100 or more units) and would thereby require greater subsampling, they are placed in the rare events universe for the rotation group in which they appear. They remain members of that rotation group for eight CPS samples, thus greatly reducing the subsampling rate.
3. The next stage in the weighting process is the adjustment for total noninterviews. For noninterview adjustment purposes, the CPS PSU's are divided into 72 clusters formed by grouping together PSU's with similar characteristics defined by the 1970 Census. The clusters are classified by geographic region, and within a region they are divided into clusters totally comprised of PSU's in SMSA's (Standard Metropolitan Statistical Areas) and those containing only non-SMSA PSU's. The clusters are further partitioned into six race-residence cells for both SMSA and non-SMSA and are applied for each of the CPS eight rotation groups.

   The weighted estimates with the duplication control and the noninterview adjustment are referred to as the unbiased estimates.
4. In order to reduce the contribution to the variance arising from the sampling of PSU's, the first stage ratio adjustment procedure is applied to records in nonself-representing PSU's. Separate ratios are used for race and residence categories of two groups of strata (SMSA and nonSMSA) for each of the four census regions. The adjustment factor is computed as the ratio of the 1970 Census population in the race-residence cell for the given cluster to the estimate of this population based on the 1970 Census population for sample PSU's in the same cluster and is applied to each of the records in the given cluster. Though a few factors are higher, a maximum factor of 1.3 is used.
5. Following the first stage ratio adjustment, the second stage ratio adjustment is applied to all sample records. The procedure attempts to bring the sample estimates into closer agreement with independent estimates of the U.S. population by age, sex, and race.

   The second stage ratio adjustment has two steps. First, a separate ratio adjustment is applied to blacks and other minority race sample persons; the basic reason for this procedure is to insure that the effect of the second stage ratio adjustment to "other races" is not weakened by the adjustment to blacks. In the next stage separate ratios are computed by sex, race (white and black and other races), and 17 age groups, giving a total of 68 cells by rotation group. The age-sex-race groups are given in Table 4 as well as the second stage adjustment computed over all rotation groups in March 1975. These factors were not the actual factors used, since they were not computed by rotation group and they assume that the intermediate adjustment was not previously applied.

Table 4. CPS Second Stage Ratio Adjustment Factors for Total Population By Age, Color and Sex 1/
ALL ROTATION GROUPS, MARCH 1975

Age	White		Black and Other Races2/	
	Male	Female	Male	Female
Total	1.04901	1.02342	1.15468	1.07532
14-15	1.01927	.99287	1.02938	1.01576
16-17	1.05079	.97006	.98710	1.10733
18-19	1.08621	1.02334	1.18278	1.14973
20-21	1.04730	1.02451	1.53855	1.18612
22-24	1.12071	1.13036	1.17701	1.07878
25-29	1.07204	1.02624	1.24781	1.06004
30-34	1.03480	1.00931	1.26153	1.13815
35-39	1.07660	1.03174	1.07273	1.13769
40-44	1.05811	1.02347	1.28741	1.10292
45-49	1.04750	1.00498	1.24003	1.03712
50-54	1.01799	1.01848	1.10833	1.07060
55-59	1.08018	1.02333	1.07344	1.02268
60-61	1.02980	.95590	.99989	1.02240
62-64	1.03221	1.07219	1.14459	.92967
65-69	1.02640	1.03032	.99706	1.05466
70-74	.96956	.98466	1.00452	.85741
75+	.98257	1.04738	1.29061	1.10894

1/ Bailar, Bailey, and Corby, 1977
2/ The factors for Black and Other Races indicate the seriousness of the undercoverage problem. These factors are not the actual adjustment factors used.

It has been shown that the ratio estimates based on age-sex-race group reduce the sampling variability of the estimates. The result of the weighting procedure is that records have weights that vary considerably.

Table 5 shows the range of weights applied to records in March 1975. It is assumed that this differential weighting will reduce both biases and variances.

B. Potential Problems with the Weighting Procedure

Some implications of the weighting procedure are as follows:

1. There is no known unbiased method of adjustment for nonresponse and undercoverage. The basic assumption is that the characteristics of the nonrespondents and missed persons are similar to those of respondents with similar demographic characteristics. An investigation of the nonrespondents was attempted in 1965 by Palmer and Jones (1967) when an intensive field followup was conducted. The results indicated that the noninterview adjustment procedures did not distort the number of employed persons. However, the results are inconclusive since less than half of the nonrespondents were interviewed. Specifically, refusals were not included in the study. It is not unreasonable to assume that the refusals have a different employment rate than the interviewed. The nature of the noninterviews has been continually changing in the last few years. In 1970 the overall Type A rate was 4.0 percent. Of these, 28 percent were not-at-homes, 23 percent were temporarily absent, 39 percent were refusals, and 9 percent were "other". In 1976 the overall Type A rate was 4.4 percent. Of these, 19 percent were not-at-homes, 17 percent were temporarily absent, 59 percent were refusals, and 5 percent were "others". The percentage of refusals is growing, and we do not know the effect of the noninterview adjustment for

Table 5. Maximum, Minimum and Average Weights for Records in 13 Relationship Categories, March 1975 1/

Relationship Category	Weights Maximum	Minimum	Average
Male head with relatives	7448.80	33.56	1645.03
Male head without relatives	8006.21	206.62	1679.78
Wife of head	7215.72	31.46	1604.91
Female head with relatives	8549.27	33.04	1621.36
Female head without relatives	8288.90	144.76	1612.22
Male child related to head	6666.70	27.12	1617.02
Female child related to head	6597.57	30.52	1551.92
Male relative (over 18)	7060.87	67.29	1695.01
Female relative (over 18)	6296.77	39.30	1625.48
Unrelated male child	6365.42	206.62	1736.01
Unrelated female child	4496.97	1153.54	1628.61
Unrelated male (over 18)	3840.59	756.86	1695.63
Unrelated female (over 18)	4369.49	991.51	1638.46

1/ Bailar, Bailey, and Corby, 1977

those who refuse although we can place at least approximate bounds on the maximum possible effect.

2. To obtain the noninterview adjustment factor the household count within the race-residence cells are tabulated by the race and residence of a designated member of the household, generally the wife of the head. However, noninterview adjustment factors are applied by the race and residence of the individual person. Thus mixed race households will not receive the same weight used in the calculation of the noninterview adjustment factor. However, mixed race households, particularly in the CPS, represent a small proportion of total households. CPS interviewers are instructed to ask race only of persons unrelated to the head; otherwise he/she records race by observation only.
3. There is a maximum factor of 4.0 used in the duplication control procedure to control the contribution of this procedure to variance, even when subsampling is at a rate greater than 1 in 4. However, this occurs very infrequently, since as previously mentioned, large USU's become a part of the rare events universe.
4. There is no evidence that the separate ratio adjustment for blacks and others which is immediately followed by the ratio adjustment for all age, sex, and race groups has a positive effect on the estimates. The factors for "other races" are highly variable.
5. The second stage ratio adjustment procedure is limited in that the census undercount is reflected in the independent estimates used to adjust the sample data during this stage of the weighting procedure. This need not be the case, however.

In summary, good measures are not available as to the impact of the weighting procedure on biases and on the effect of errors occurring in earlier stages of the survey.

VIII. Conclusion

This paper has attempted to present in terms of an error profile the potential sources of nonsampling error in the Current Population Survey. Because nonsampling errors have historically been more difficult to measure than sampling error, less is known about their effect on the data. This is unfortunate since for some statistics they may dominate the mean square error. As indicated in the paper, studies have been performed on the Current Population Survey that give some indication of nonsampling error on specific survey operations, e.g. the Methods Test and the CPS/FOSDIC Study. In addition, a reinterview study is conducted each month to measure response bias. However, the reinterview survey has its limitations, some of which may be difficult to overcome, and many of the studies discussed in the paper were limited. As such, there are large gaps in our knowledge of nonsampling error. It is hoped that this error profile of a major complex survey such as the Current Population Survey for which accuracy is a continuing concern, will stimulate the accumulation of more knowledge of this subject in all surveys.

Acknowledgements

This paper was written as a result of the work of the Subcommittee on Nonsampling Errors of the Federal Committee on Statistical Methodology, and the authors wish to thank the members of the Subcommittee for their suggestions, comments, and support. However, it represents the compilation of a great deal of work by Bureau staff members. The authors wish to acknowledge those contributions. They are appreciative of all the help they received from numerous persons in the Bureau who called to their attention unpublished Bureau studies and assisted in their interpretation. They wish to particularly thank persons in the Statistical Methods Division, the Engineering Division, and the Demographic Surveys Division, for making available requested information, suggesting sources of additional information, and for their many helpful comments and suggestions.

References

[1] Baer, L., "Permit Status of TAR ED's in the Redesign," Bureau of the Census Memorandum to M. Boisen, January 5, 1973.

[2] Bailar, B.A., Bailey, L., and Corby, C., "A Comparison of Some Adjustment and Weighting Procedures for Survey Data," Unpublished paper presented at Sampling Symposium, Chapel Hill, North Carolina, 1977.

[3] Boisen, M., "Determination of Type of Segment for Tape Address Register (TAR) ED's Which Are in Non-permit Areas," Bureau of the Census Memorandum to J. Waksberg, July 26, 1971.

[4] Boisen, M., "Final Report on CPS/FOSDIC Experiment," Bureau of the Census Memorandum to D. Levine, June 27, 1975.

[5] Deighton, R., "Methods Test Report: Some Results of Experimentation with Self-Response Interviewing Procedures, February, 1965 - June, 1966," Bureau of the Census Memorandum, February 28, 1967.

[6] Hanson, R.H., Statistical Methods in the Current Population Survey, Draft of the Revision of Technical Paper No. 7, 1976.

[7] Hirschberg, D., Yuskavage, R., and Scheuren, F., "The Impact on Personal and Family Income of Adjusting the Current Population Survey for Undercoverage," Unpublished paper presented at the Annual Meetings of the American Statistical Association, Chicago, 1977.

[8] Jablin, C., Unpublished Monthly Report of the Methods, Procedures, and Quality Control Branch of the

Demographic Surveys Division, Bureau of the Census, 1977.

[9] Johnston, D.F. and Wetzel, J.R., Effect of the Census Undercount on Labor Force Estimates, Special Labor Force Report No. 105, Bureau of Labor Statistics, March 1969.

[10] Jones, C. and Blass, R., "Population Undercoverage Estimates from the CPS-Census Match," Bureau of the Census Memorandum to H. Nisselson, January 17, 1975.

[11] MacKenzie, W., Documentation of Annual Housing Survey Coverage Improvement Program, Bureau of the Census, 1977.

[12] McPherson, J. L. and Volk, M., "FOSDIC Microfilm Problems and their Solutions," Unpublished Paper presented at the Eleventh Annual Convention of the National Microfilm Association, Washington, D.C., April 25-27, 1962.

[13] Moye, D., "CPS Reinterview Results from the Listing Check, Check of Noninterview Classifications, and the Household Composition Check for 1975," Bureau of the Census Memorandum, May 21, 1976.

[14] Palmer, S., "On the Character and Influence of Nonresponse in the Current Population Survey," Proceedings of the Social Statistics Section, American Statistical Association, 1967, pp 73-80.

[15] Schreiner, I., "CPS Reinterview Results from the Listing Check, Check of Noninterview Classifications, and the Household Composition Check for 1976," Bureau of the Census Memorandum, April 27, 1977.

[16] U.S. Bureau of the Census, Census of Population and Housing: 1970 Evaluation and Research Program PHC (E)-4, Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis by J.S. Siegal, U.S. Government Printing Office, Washington, D.C., 1973.

[17] Williams, L. E., "Methods Test Phase III: First Report on the Accuracy of Retrospective Interviewing and Effects of Nonself-Response on Labor Force Status," Bureau of the Census Memorandum to W.M. Perkins, June 24, 1969.

AN ERROR PROFILE: EMPLOYMENT AS MEASURED BY THE CURRENT EMPLOYMENT STATISTICS PROGRAM

Lillian H. Madow, Bureau of Labor Statistics of the U. S. Department of Labor

I. Introduction

An error profile is a vector, each component of which corresponds to an aspect of the survey process that may lead to errors in the data. It is desirable to have sampling error and a measurement of overall error among the components of the error profile. The components may be overlapping or not, correlated or not. Some of the components of an error profile may be scalars, for example, estimates of variance components; others may be vectors or matrices, i.e. tables of characteristics of sample and population. The purpose of this study is to present steps towards an error profile for the national employment estimates of the Current Employment Statistics (CES) Program of the Bureau of Labor Statistics (BLS).

II. Specifications Met by the Survey

1. Purpose. The CES program provides monthly estimates of employment, and hours and earnings of persons on the payrolls of nonagricultural establishments including government, by detailed industry. Estimates are published for the Nation, States [1/] and local areas. This paper is concerned only with the estimates of employment for the Nation and for eight major industry divisions.

2. Publication Dates. Preliminary estimates for at least the Nation and for 8 industry divisions, are published in a press release issued the third Friday after the week including the 12th of the month. They are also published with more industry detail about two weeks later in the BLS monthly periodical Employment and Earnings (E&E). Estimates are published for over 400 industries, or aggregations of industries, in E&E, in each of the two following months. The three sets of estimates are often called the first, second and third closing estimates.

3. Relative Error. The relative error of the estimate of National month-to-month change, the ratio of the current month's estimate of employment to the preceding month's estimate of employment, is between 0.1 and 0.2 percent.

4. Administration. The CES is a joint Federal-State cooperative data collection and processing program. The States prepare the CES estimates for the States and local areas. BLS-Washington prepares the National estimates. The BLS Regional Offices provide guidance and technical assistance to the States.

III. Concepts: Establishment, Employment, Industry

The concepts of employment and industry are fundamental in the CES program, because estimates are produced of employment by industry.

1. Establishment. An establishment is defined to be an economic unit such as a farm, mine, factory, or store which produces goods or provides services; it is usually at a single physical location and engaged in one type of economic activity. If more than one type of economic activity is performed at a single location, each activity is treated as a separate establishment, provided that:

a. No one industry description, at the level of industry detail considered, includes the combined activities;

b. The employment in each such activity is significant;

c. Reports can be prepared on the number of employees, their wages and salaries, sale of receipts and other establishment type data;

d. The enterprise owning the establishment is willing to provide the reports on employment and other information for each of the establishments.

Thus, an establishment is not necessarily the same as a business or company; these may consist of several establishments.

The unit that reports information in the CES program is called a reporting unit. Often, the reporting unit is an establishment. Sometimes, a reporting unit consists of several establishments, e.g. a chain store may provide a single report for all of its establishments in a county. Sometimes, as in the transportation or public utility businesses, the concept of the establishment being at one location does not apply.

2. Employment. Establishments report the number of employees on their payrolls who receive pay for any part of the pay period including the 12th of each month. For most establishments, this pay period is a week but an establishment reports for whatever pay period it actually uses, bi-weekly, monthly or other. CES estimates are also prepared for women employees and for production and nonsupervisory employees, but this study only considers all employees. A person will be counted as many times as the number of payrolls on which he is listed for the reference period, whether because of holding more than one job or because of changing jobs during the specified pay period. Proprietors, the self-employed, unpaid volunteer or family workers, farm workers, and domestic workers in households are excluded according to the above definition, but employees at all levels are included, e.g. executives of corporations. Government employment covers only civilian employment; military personnel are excluded.

There is no requirement that a minimum number of hours be worked during the pay period; the only requirement is that the person be on the payroll and be paid.

3. Industry. Industries are classified according to the Office of Management and Budget (OMB) Standard Industrial Classification (SIC) [2/] code, with a 1,2,3, or 4-digit code -- the higher the digit, the more detailed is the classification. The higher digits are subsets of the lower digits and can be aggregated to form different levels of industry groupings. To facilitate classification by industry, establishments provide information on their principal products or activities and the percent of sales value or receipts resulting from each of these products.

IV. Estimation

1. General Description. CES estimates of employment are first computed for 846 estimating cells -- an estimating cell consists of all establishments in an industry defined by a 3 or 4-digit SIC code; some of the industries are further subdivided by region and/or size of establishment as measured by employment. Then, the estimates of total employment for the 846 estimating cells are summed to provide estimates for larger industry groupings.

For each estimating cell, the CES estimate of employment is a product of three terms:

- A benchmark, B. The benchmark is a relatively complete count of employment computed for March of every year, with some exceptions, but not available for about 18 months after the reference month, March;

- A product of link relatives, L. The link relative for a specified month is the ratio of total employment in that month to total employment in the preceding month for establishments reporting in both months. (In the actual estimation process, as discussed in Section V, the estimator may be more complex.);

- A power of an adjustment factor, F. The adjustment factor estimates the effects of births, deaths and other "persistent" sources of bias on employment.

2. Estimators. An estimator, E_{ik}, of employment for the k^{th} month after the last benchmark available at the time the estimator is computed has the form:

$$E_{ik} = B_{io} L_{i1} \cdots L_{ik} F_i^k \qquad (1)$$

where i denotes the estimating cell, B_{io} is the benchmark, L_{ij} is the link relative for month j, j = 1, ...,k, and

$$L_{ij} = \frac{Y_{i,j,j}}{Y_{i,j,j-1}},$$

where $Y_{i,j,h}$ is the total employment in cell i in month h (h=j, j-1), after the benchmark for establishments reporting in month j after the benchmark, and F_i is the adjustment factor. Thus, the first subscript identifies the cell, the second subscript identifies the month for which the link relative is computed and the third subscript identifies the month of the data summed.

The subscript, i, will now be omitted for convenience. Let $B_{-\alpha}$, α= 1, 2, ..., 6, be the last 6 benchmarks at intervals of one year prior to B_0.

In order to state how F is calculated, let us define $E'_{-\alpha}$ by

$$E'_{-\alpha} = B_{-(\alpha+1)} L_{\alpha 1} \cdots L_{\alpha 12}, \ \alpha = 1, \ldots, 5 \qquad (2)$$

where $L_{\alpha 1}, \ldots, L_{\alpha 12}$ are the link relatives for the 12 months following the month of $B_{-(\alpha+1)}$. Then the adjustment factor, F, is

$$F = 1 + \frac{1}{60} \sum_{\alpha=1}^{5} \frac{B_{-\alpha} - E'_{-\alpha}}{B_{-\alpha}} \qquad (3)$$

Usually, values of F are close to 1, ranging from 1.000 to 1.004, or in a few cases a little larger, but more often, no greater than 1.002.

As expression (3) shows, any source of bias that is persistent over the 5 benchmark comparisons included in (3) will affect the value of F.

The form of the estimator given in (1) and the fact that the benchmark and link relatives use the same concepts of employment imply that current estimates of employment are extrapolations of the benchmark based on the link relatives and adjustment factor.

The relative change in employment from month k-1 to month k is estimated by

$$\frac{E_k - E_{k-1}}{E_{k-1}} = L_k F - 1 \qquad (4)$$

Thus, estimates of relative month-to-month change are independent of the last available benchmark and depend only on the current link relative and the adjustment factor, F. Since F is usually small and has no cumulative effect in a one month period, relative month-to-month change depends primarily on the current link relative.

Adjustment factors are computed by BLS with each new benchmark, for selected 2 and 3-digit industries. Thus, more than one industry may have the same adjustment factor.

In order to discuss the formulae for first, second and third closing estimators, it is desirable to define first, second and third closing dates. By reference week for a given month is meant the calendar week containing the 12th of that month. All three closing dates occur on a Monday. The first closing is the third Monday after the reference week. The second and third closings occur at three week intervals after the first closing.

If E_k is a first closing estimator, then L_k is computed for establishments whose data are received in BLS by the first closing date for month k, and L_{k-1} is computed for establishments whose data are received in BLS by the second closing date for month k-1. Link relatives for months 1, 2, ..., k-2 are third closing link relatives.

If E_k is a second closing estimator, then L_k is computed for establishments whose data are received in BLS by the second closing date for month k. Link relatives for earlier months are third closing link relatives.

If E_k is a third closing estimator, then all k link relatives are third closing link relatives.

As mentioned earlier, link relatives are computed for establishments providing data in both the month of the link relative and in the preceding month.

BLS computes estimates for each of the 846 estimating cells. In general, estimating cells include several of the strata used in selecting the sample. BLS does not, however, use sampling weights for responses in the different strata within an estimating cell.

Although not discussed in detail here, BLS uses several means of detecting outliers and reducing their effects on the estimators. These have the effect of smoothing month-to-month changes.

V. Steps in the Production of CES Estimates

Chart 1 shows the major steps in the production of National CES employment estimates. It provides an overview of the CES design.

One major data source of the CES estimates is the Unemployment Insurance (UI) file. Each quarter, mandatory tax reports containing monthly data on employment and quarterly data on wages are submitted to the States by over 4,000,000 reporting units subject to State UI laws. The UI data are supplemented by data from the Civil Service Commission (CSC), the Census of State and Local Governments, the Interstate Commerce Commission (ICC), and other sources.

The UI program provides:

- The ES-202 report which is a summary of the UI tax reports. Each quarter, the States edit and tabulate the UI data by industry code (in the first quarter, by industry code and size of establishment). These tabulations give the distribution of establishments and payroll employees by industry and size class. They essentially constitute the ES-202 first quarter reports, which are due at BLS-Washington 6 months after the quarter ends.

The ES-202 reports are used to compute benchmarks and to determine the size of the incremental sample.

- The Unemployment Insurance Address File (UIAF), which is a listing of establishment identification and various characteristics, including number of employees and industry code. Each State prepares this list annually from the UI tax reports. In 1978, UIAF will include over 98.5 percent of all establishments.

The UIAF provides the frame for the selection of the incremental sample selected each year to update the 790 Survey.

The other major date source is the BLS 790 Survey, a national survey, called the 790, because of the form number.

The BLS 790 Survey is a voluntary, mail, monthly survey in which approximately 160,000 establishments report each month on total employment, and the employment of women and production workers as well as the hours and payrolls of production workers.

The 790 Survey is used to calculate the link relatives.

The last major revision of the 790 Survey occurred in 1963, when an improved design was introduced. This design was a stratified random sample of establishments from each of over 400 industries, based on the then existing UIAF, supplemented by samples from industries not included in the UIAF.

Within each industry in the UIAF, the stratification was by size of establishment and effectively by State since each State used National sampling ratios to select its sample from the State UIAF.

In almost all industries, all establishments having 250 or more employees constituted the certainty stratum and were designated for the sample. In some industries, all establishments having 100 or more employees constituted the certainty stratum. The sampling ratio from each of the other strata was proportional to the average size of establishment in the stratum, as determined by the ES-202 tabulation for the first quarter in 1963.

If the coefficient of variation of size of establishment was constant for all non-certainty strata for a given industry, the stratified sampling design used optimum allocation for the non-certainty strata.

Exceptions to the stratified design occur for industries not listed in the UIAF. It may be that neither an establishment list (including establishment sizes) nor the equivalent of the ES-202 tabulation is available.

Incremental Sample. Each year, after the ES-202 listings for the first quarter become available, the States update the sample by selecting an incremental sample. They are expected to compute the desired size of sample within each stratum by multiplying the sampling ratio by the number of establishments in a size class in the ES-202 report for the State. Then the States are to select at random, or systematically, from the corresponding stratum of the State UIAF, a sample consisting of the number of establishments equal to the difference between the expected number and the current actual sample size.

The benchmark is computed annually (with some exceptions) from the ES-202, supplemented for industries not covered by UI, and modified by industry classification information from the 790 Survey.

The link relatives are computed monthly from the 790 Survey.

The adjustment factor is computed annually from benchmarks and link relatives for 5 periods preceding the last available benchmark.

VI. The Error Profile

1. An Approach. The number of possible components of an error profile of an estimate, which depends in whole or in part on a survey of respondents, is very large. It seems reasonable, therefore, to begin by identifying the major sources of error of the estimate and to relate the source to the components of error. In the development of the CES error profile, it has been convenient to organize the profile and identify the components according to the "paths" between steps in Chart 1. First, however, relevance and concepts are considered; these remain sources of error even if sampling is not used.

2. Relevance and Concepts.

a. Relevance. The components of an error profile measuring relevance -- roughly, how much the survey information (even if "true") differs from what is desired, may well vary from user to user. Further, the means of approximating such components are rarely set down. Early in planning, judgments are made on what is reasonably practicable and from then on

discussions are in terms of the desired information. In a continuing survey, perhaps the best means of studying relevance at any given time is to consider with what objectives the analysts are transforming or adjusting the estimates, and what they say about the estimates.

b. Concepts. To what extent do the definitions of concepts inherently define random variables rather than constants? Is the schedule in agreement with the desired concept? Are the reported responses those called for by the schedule? If the same information is obtained from two or more sources, perhaps at different times, as in a ratio or regression estimator, are the concepts the same for the different sources? What are their measurement errors?

The UI tax reports and 790 survey are based on the same concepts of employment, establishments and SIC codes. More instructions are given for the 790 survey than the UI tax reports. The forms used by UI vary from State to State, although the same employment question is asked. The last reported study was by Young and Goldstein.[3] It showed that the 790 schedules were filled in almost exlusively from payroll records, and that the net effects of incorrect reporting were very small.

The 790 assignments of SIC codes are compared annually with UI assigned SIC codes. It will be found in the discussion of the benchmark below that, on the level of industry divisions, there is apparently little difference between UI and the 790 survey in the assignment of SIC codes.

Let us return to the chart and consider error components for the major steps in obtaining CES estimates with the branches leading to the three factors on which the estimates depend: benchmark, B, link relatives, L, and adjustment factor, F.

3. From UI to Benchmark.

a. Imputation of UI Tax Reports. States summarize the tax reports for each quarter, containing monthly data on employment in a report called the ES-202 report. Three months after the end of the quarter, the State Employment Security Agency (SESA), imputes for establishments whose reports have not arrived. Imputation accounts for from 2 to 10 percent of the establishment reports but no more than one percent of employment. At present, there are general guidelines for imputing.

b. Benchmark. Benchmarks are computed almost every year primarily from the ES-202 for the first quarter of the year. The computation of the benchmarks begins with ES-202 reports for the 50 States and the District of Columbia. SIC codes of establishments in the 790 survey and of UI tax reporting units are compared; the more detailed specification of establishments with different SIC codes is adopted and ES-202 data are modified by transferring employment in accordance with changes in SIC's.

Then, data on total employment for SIC's not in the ES-202 report are obtained and added to the modified ES-202 data to obtain the benchmarks.

Table A shows the steps from the ES-202 reports (after they are summarized in BLS) to the benchmark. The most important step is that of adding employment for the SIC's that do not have full UI coverage. These magnitudes are shown by industry division in Table A, and the details and sources of the estimates are given in Table B.

The small changes resulting from SIC assignments are shown in the third column of Table A. The column headed ES-202 will differ slightly from data published in the BLS quarterly periodical, Employment and Wages (E&W). The published table in E&W includes the 50 States, the District of Columbia, and Puerto Rico. The column headed ES-202 includes only the 50 States and the District of Columbia, since National CES estimates do not include Puerto Rico.

Table B shows the SIC's and the estimated employment for those industries in which the UI is supplemented. The sources from which the supplementary data were obtained are the Interstate Commerce Commission (ICC); County Business Patterns (CBP), an annual publication of the Census Bureau based on data obtained from the Social Security Administration; The American Hospital Association (AHA); the Center for Education Statistics and the Office of Education of the Department of Health, Education and Welfare (HEW).

c. Completeness of Benchmark. How "reasonably complete" a count is the CES benchmark? Some establishments may not file tax reports with the States, or may file their initial reports late. Also, the estimates of employment for the SIC's not covered by UI are not necessarily precise. One indication of the completeness of the CES is how it relates to the Current Population Survey (CPS) estimate of employment. One study by Green [4] and another by Korns[5] compared the CES estimate of jobs with the CPS estimate of employment converted to an estimate of jobs. The conversion primarily consists of adjusting for the number of jobs held by persons with more than one job and for the number of jobs held by persons not counted in the Census. The latter estimates depend on hypotheses concerning the employment characteristics of those not counted in the Census. If the effects of the Census undercount are ignored, then the CES estimate of jobs exceed the adjusted CPS estimate. If one accepts the Green and Korns estimates of undercount, the CES estimates are less than the adjusted CPS estimates by about 1.5 percent. If one accepts the Johnson and Wetzel [6] assumptions (in a Bureau of the Census study of employment and unemployment effects of the undercount), the CPS adjusted estimate might be one percent greater than the CES estimate of jobs. In view of the differences in concept, samples, and data collection procedures, as well as the assumptions made in the undercount studies, the conclusion seems to be that CES and adjusted CPS estimates do not differ importantly. Also, the benchmark and revised CES estimates differ by about 0.1 to 0.2 percent. The conclusion is that the benchmark is a reasonably complete count.

4. From UI to 790 Survey and the Link Relatives.

a. The Frame. Every year, each State is to use its Unemployment Insurance Address File (UIAF) as the frame from which to select incremental samples in

order to maintain the 790 sample. For industries not having UI coverage, the States select their own frames, and select and maintain their own samples. Currently in progress is a special Survey of the States in which the State agencies will report on the procedures used and the problems encountered in maintaining the 790 Survey.

The completeness of the UI portion of the total frame is best indicated by the fact that the establishments listed in the UIAF account for about 97 percent of all employment in private nonagricultural industries, all Federal employment, 80 percent of State Government employment and 15 percent of local government employment. The completeness of the frames used for most of the industries not covered by UI is not known. However, beginning with the first quarter of 1978, UI, and therefore, UIAF coverage will include about 98.5 percent of all employment in nonagricultural establishments, primarily because all State and local governments will have UI coverage.

b. Comparison of 790 Sample with Universe and Potential Sample. Table C shows a comparison of the actual sample for Marc, 1974, with the potential size of sample, the latter being obtained by applying the sampling ratios for individual industries to the tables of employment by industry and size class in the ES-202 report, with some adjustments for industries not covered. Table C is presented here to show the relationship of the employment in establishments reporting in the 790 survey to both the universe population and the potential sample size, if there is neither refusal nor no nonresponse. The table displays the large size of the 790 sample.

c. Processing the 790 Schedules. As illustrations of the numbers of schedules processed each month: in April, 1977, of the 159,843 schedules entering the editing and screening module (which includes matching), 8,468 were not used in estimation, either because the data were rejected during editing and screening or because there were no data for the establishment in the preceding month. (The latter data are required by the use of link relatives.) Thus, the estimates were based on 151,374 schedules, or 94.7 percent of those entering the editing and screening module. The corresponding data for May were 156,613, 6810, and 149,803, or 95.7 percent of those entering the editing and screening module. In both months, the number of schedules used in making estimates was about 85 percent of the National Registry of active reporters which lists about 184,000 establishments. However, the value, 85 percent, is a response rate only for the National Registry. In a voluntary continuing survey of establishments, refusals and "dropouts" occur. The National Registry includes only "active reporters". The comparison of estimates and benchmarks presented later includes the net effects of nonresponse and the selection of respondents.

During the data processing operations, including estimation, listings are prepared of establishments and estimating cells that fail various editing and screening tests including comparisons with past data for establishments or cells. It would be useful to have summary tables prepared in addition to the listings, and also to learn what would have been the estimates, if editing or screening or reviewing estimates by estimating cell were not done.

At the conclusion of the monthly data processing cycle for a given closing estimator, the link relatives have been calculated and the estimation formulae (Section V) are applied.

5. Comparisons of Estimates and Benchmarks. Benchmarks are available about 18 months after the benchmark month. At that time, the estimates for the benchmark month and the benchmark can be compared.

Four estimates for the benchmark month are computed. The first, second and third closing estimators, here denoted by E_1, E_2, and E_3, are computed using the last available benchmark at the time of computation, usually that of 24 months previous to the benchmark month. About 6 months after a specified benchmark month, the benchmark for the preceding year, 12 months prior to the specified benchmark month, usually becomes available. The fourth estimator, E_r, is computed from the newly available preceding year's benchmark, the link relatives for the following 12 months and the adjustment factor. Comparisons of E_1, E_2, E_3, E_r with B, the benchmark for the same month as the four estimates, provide one basis for evaluating CES estimates. Another basis is how useful these data are in analysis; this second basis is not discussed here.

Let us review the major sources of error to identify those whose net effect is included in the benchmark comparison.

The concepts of employment, establishment and SIC used in the CES estimates and the benchmark are the same. There may be differences in implementation, but until this question is studied, the comparisons made below of estimates and benchmark cannot now be said to include errors attributable to concepts and their implementation.

The comparison between estimate and benchmark does include the effects of both respondent selection and nonresponse, but does not reflect the previously discussed possible small undercount in the benchmark since the frame used for the 790 survey is a list of the establishments whose data are the major part of the benchmark.

Data processing to obtain the ES-202 State Reports from the UI tax reports and the first editing of 790 schedules are performed by the States using instructions prepared by BLS-Washington. Data processing is performed by BLS-Washington for the benchmark and estimates, using and supplementing the State ES-202 reports, the 790 schedules and the adjustment factors. Thus, the comparison between benchmark and estimates may reflect some differences due to data processing.

In Table D, relative differences between the estimates and benchmarks are presented as well as relative differences between first and third closing estimates for the same months.

It would be easy [7/] to compute summary statistics from Table D. However, the number of years is only three and the summary statistics might mask the essential close agreement not only for total employment but also for the 8 industry divisions.

Detailed data are shown for only the three years since the last major increase in UI coverage. The next major increase in UI coverage will affect primarily State and local governments. The comparisions for the private sector, at least, may be expected to be stable.

The conclusions from Table D are:

a. Agreement between first and third closings is reasonably good.

b. Differences between third closing and benchmark measure the error in the level of the estimate. It is difficult to generalize concerning the current level of the mean square error, since the only comparison for a 24-month period with the present level of coverage is the 1975 comparison which shows small mean square errors, except for mining (which has relatively small employment) and government (which should improve beginning in 1978 with the increase in UI coverage).

In the formula for the relative mean square error of the third closing estimate of level, k months after the last available benchmark, one term is the product of k^2 and the relative mean square error of the adjustment factor. The factor k^2 will lead to a large relative mean square error of the estimate of level resulting from this term, if k is large enough. The data of Table D confirm this.

c. The squared relative errors, $(E_p-B)^2/B^2$, provide upper bounds for the ratios of current month estimated employment to preceding month estimated employment.

VII. Summary and Conclusions

An error profile may contribute to the achievement of several possible goals:

1. Improvement of analysis through the measurement of overall error;

2. Optimum allocation of resources among the parts of a program, e.g. for a given budget, to allocate resources among the different parts of a program to minimize overall error, or for a given overall error, to allocate resources among the different parts of a program to minimize cost;

3. Understanding the limits of possible achievement by spending more money without changing design since nonsampling biases may not tend to zero as the size of sample increases;

4. Identifying aspects of the survey on which efforts should, if practicable, be made to reduce the contributions to the mean square error arising from those aspects;

5. In continuing surveys, identifying survey aspects, where deterioration is occurring and remedial action is needed;

6. Providing to the designers of computer programs, a list of outputs that will be useful in routinely measuring error components arising in the computer process, without special studies;

7. To make possible improved analysis of relationships among the underlying "true values";

8. To study cost effectiveness.

To achieve these goals may be costly except, perhaps, for the computer requirements in item 6 (if they are developed early enough), but the benefits of achieving the goals will be great. For large and continuing surveys (and also for some smaller surveys), items, in addition to sampling and overall errors, that would justify continuing measurement efforts are:

1. Concepts;

2. Changes in the population and frame;

3. Completeness of frame;

4. Data collection procedures, including
   a. Agreement to participate;
   b. Dropouts, permanent or temporary;
   c. Current and cumulative response rates; and
   d. Response errors due to data collection;

5. Any imputation process whether explicit, e.g. substitution of another schedule, use of past data for element, adjusted or not, or implicit, e.g. weighting procedure - and whether for nonresponse, missing data or "outliers";

6. Steps in data processing, including estimation, both for
   a. Correctness of processing steps, e.g. card punching, and
   b. Detection of data errors or outliers, e.g. editing and screening;

7. Implications of analysis requirements for the accuracy of the survey estimates;

8. Cost-effectiveness of the survey.

Many possible error components have been discussed in the preceding pages, but few could be estimated from information currently available. Much of the necessary data already exists and is used in the CES system. A research program has been developed to provide improved measurement of an error profile.

Footnotes

1/ In this report, the meaning of States includes the 50 States and the District of Columbia; the CES Program is also conducted in the Commonwealth of Puerto Rico, but those estimates are not included in the U.S. National estimates.

2/ Standard Industrial Classification Manual, Office of Management and Budget, Executive Office of the President. The 1967 edition of the Manual is currently being used in the CES program. CES estimates based on the 1972 edition of the Manual are expected to be published in the Fall of 1978.

3/ Young, Dudley E. and Goldstein, Sidney, "The BLS Employment Series and Manufacturing Reporting Practices", Monthly Labor Review, November, 1957, pp. 1367-1371.

4/ Green, Gloria P., "Comparing Employment Estimates from Household and Payroll Surveys", Monthly Labor Review, December, 1969.

5/ Preliminary research by Alexander Korns, Bureau of Economic Analysis(BEA), using 1974 data.

6/ This is discussed in Brooks, C.A. and Bailar, B.A., "An Error Profile: Employment as Measured by the Current Population Survey", presented at the American Statistical Association meeting in Chicago, August 15, 1976.

7/ Summaries are included in section "Explanatory Notes" published monthly in E&E.

Acknowledgment

This paper was written as part of the work of the Subcommittee on Nonsampling Errors of the Federal Committee on Statistical Methodology, Statistical Policy Division, Office of Management and Budget. I should like to thank the members of the Subcommittee for their suggestions, comments, and support. Also, I should like to thank the many people at BLS for all the help they gave; in particular, the members of the Office of Employment Structure and Trends, headed by Dudley E. Young, and the Office of Survey Design, headed by Barbara A. Boyes.

Chart 1. Current Employment Statistics (CES) Estimates of Employment.

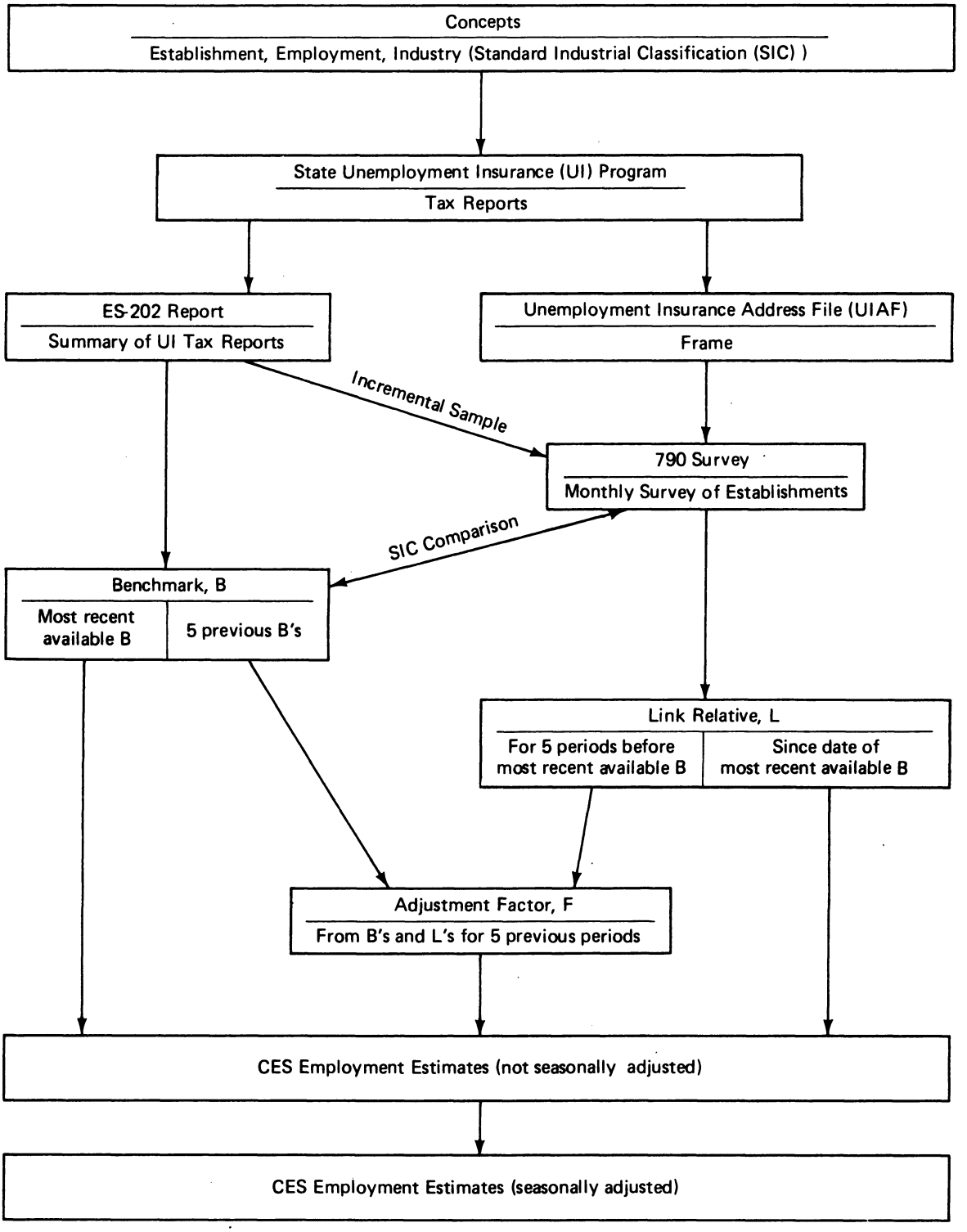


Table A. Employment in Nonagricultural Establishment, From ES-202 to Benchmark, by Industry Division, March, 1974.
(thousands of employees)

Industry Division	ES-202[1] summary	Changes due to changes in SIC	Other	March, 1974 Benchmark
	(1)	(2)	(3)	(4)
Total private [2]	61,144	0.0	2017.8	63,162
Mining	699	-1.1	0.0	668
Contract Construction	3,760	1.5	0.0	3,762
Manufacturing	19,973	13.9	0.0	19,987
Transportation and public utilities	4,091	-0.6	578.5[4]	4,669
Wholesale and Retail Trade	16,566	-2.1	0.0	16,564
Finance, insurance and Real Estate	4,062	1.3	103.9	4,167
Services	12,023	-12.9	1335.4	13,345

1/ There are small differences between the column headed ES-202 and the data published in BLS Employment and Wages, (E&W) First Quarter, 1974, due primarily to the fact that Puerto Rico is included in the tabulations published in E&W but not in the ES-202 Summary.

2/ Differs from published E&W because SIC 99 (nonclassifiable establishments) and SIC's 07-09 (Agricultural Services, Forestry and Fisheries) are included in Services in CES but are not in Services in E&W.

3/ Differs from published E&W due to exclusion of SIC 01 (commercial farms) in this table.

4/ Includes 573.9 for SIC 40, Railroads (a complete count), for comparability with benchmark.

Table B.* Adjustments in Employment for ES-202 Coverage Exclusions, March, 1974

Category exempt from UI coverage	SIC	Benchmark March 1974	Benchmark Source
1. Trucking companies owned by RR	421,2	200	ICC
2. Railroad car loan companies	47	4,400	ICC
3. Nonoffice insurance salesmen	631	75,000	CBP
4. Nonoffice insurance salesmen	633	13,400	CBP
5. Nonoffice insurance salesmen	635,6,9	1,500	CBP
6. Religious trusts	67	14,000	CBP
7. Private hospitals	806	93,400	AHA
8. Private elementary and secondary schools	821	224,000	Various
9. Private Colleges and universities	822	155,600	HEW
10. Other schools & educational services	823,4,9	29,300	CBP
11. Religious organizations	866	825,100	BLS[1]
12. Nonprofit organizations with less than 4 employees		8,000	CBP
13. Total adjustments (Sum 1-12)		1,443,900	--
14. Railroad transportation [2]	40	573,900	ICC
15. Federal Government	91	2,691,000	CSC
16. State & Local Government [3]	92,93	11,589,000	Census
17. UI-Covered Private industries	-	61,144,200	ES-202
18. Total Benchmark (Sum of 13-17)	-	77,442,000	-

* Memorandum: Carol M. Utter to John Tucker, August 28, 1975, entitled "march 1974 Benchmark Adjustment," Table 6.

1/ Based on Council of Churches data plus others for 1974.

2/ Covered by Railroad Retirement Board.

3/ UI-covered partially; UI will cover almost completely in January, 1978.

Table C. Actual and Potential Samples, March 1974*

Industry Division	3/74 Benchmark	Actual BLS Sample[1/] Employees (thousands)	Actual BLS Sample[1/] Percent of Benchmark	Potential Sample[2/] Employees (thousands)	Potential Sample[2/] Percent of Benchmark
	(1)	(2)	(3)	(4)	(5)
Total[3/]	77,155	33,613	41	43,191	56
Mining	668	307	46	423	63
Contract Construction	3,762	771	20	1,544	41
Manufacturing	19,987	11,821	59	14,824	74
Railroads	574	537	94	537	94
Other transportation and utilities	4,095	2,181	53	3,576	87
Wholesale and retail trade	16,564	3,050	18	6,145	37
Finance, insurance and real estate	4,167	1,507	36	2,004	48
Serivces [3/]	13,058	2,716	21	5,415	41
Government: Federal	2,691	2,691	100	2,691	100
State & Local	11,589	6,032	52	6,032	52

*Based on a letter from M.S. Raff to N. Frumkin, Nov. 2, 1976.

1/ As reported in Table H of E&W, except as modified by footnote 3.
2/ Expected number if BLS sampling ratios were fully implemented without nonresponse.
3/ Omits service employment in agriculture, forestry, fisheries, and unclassifiable establishments (SIC 07,08,09,99).

Table D. Relative Differences[1/]: Employment Estimates and Benchmarks, March 1973, 1974, 1975.
(in percent)

Industry	First and third Closings 1973	1974	1975	Third Closing[2/] and Benchmark 1973	1974	1975	Revised Estimate and Benchmark 1973[3/]	1974	1975
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TOTAL	0.0	-0.2	0.0	-1.6	-1.7	0.1		-0.1	0.2
Mining	0.2	-0.5	0.1	-3.5	-3.9	-5.7		-3.0	-1.9
Contract Construction	-0.1	-0.4	0.1	-9.6	-9.5	0.9		0.6	0.3
Manufacturing	0.0	-0.3	0.0	-1.1	-1.3	0.1		-0.1	0.3
Transportation and Public Utilities	-0.2	-0.1	0.4	-0.7	-0.7	-0.4		0.0	-0.6
Wholesale and Retail Trade	0.1	-0.1	0.0	-2.2	-2.3	-0.3		0.1	-0.2
Finance, Insurance and Real Estate	0.2	-0.1	0.3	-0.5	-1.6	-0.2		-1.1	0.9
Services	-0.1	-0.1	0.1	-0.9	-0.5	-0.7		-0.7	-0.1
Government	-0.2	-0.2	0.0	-0.5	0.0	1.9		0.5	0.8

1/ [(earlier date - later date)/later date] · 100

2/ The third closing for March, 1973 is based on the March, 1971 benchmark, prior to the increase in UI coverage in 1972; The third closing for March 1974 is based on the March, 1971 benchmark, since there was no benchmark in March, 1972; The third closing for March, 1975 is based on the March, 1973 benchmark.

3/ The revised estimate is not available for 1973, since the 1972 benchmark was not computed.

COMMENTS ON ERROR PROFILE PAPERS BY T. DALENIUS: C.A. BROOKS AND B.A. BAILAR: L. MADOW

I. R. Savage, Yale University

An Error Profile is a systematic and comprehensive review of survey operations which calls for the measurement of the differences between what is done and what is ideal. (A paraphrase of TD)

From the examples at hand (BB and LM) we are not sure this work can be done. The evidence is not strong that a systematic or comprehensive effort is needed.

It is most doubtful that all the sizes of sources of error can be measured (TD) but maybe an expert could locate the major sources of error; presumably that is what BB and LM attempted. Notice there are several ways of measuring sizes of error sources: (1) For a particular data set compare with benchmarks (LM); (2) By studying the process and estimating the mean, variance, etc., associated with source of error.

The consumer of statistics cannot be interested in the error profile -- his concern is with expected losses from potential decisions. Total error will often be a useful measure for that purpose. The statistician likes the profile because it indicates good places for him to increase effort. The public, including scholars, want the profiles so that they can understand the work of the data producers.

My casual reading of the papers has not given me a good idea of the nature of the camel, although TD assures us the gnat is sampling error. I conjecture the camel is the lack of coordination between the producer and consumer of statistics. Methodology and resources are lacking to:

(1) Measure the utility of imperfect data in policy making or administrative action.

(2) Communicate data needs to data producers.

Although it is not always possible to study all aspects of a problem and the detailed technical work must be done, practical statistical programs should be directed at real problems. Before substantial expenditures are made we should attempt to make sure we will be analyzing the camel rather than another gnat.

SOCIAL STATISTICS IN 2000

Philip M. Hauser, University of Chicago

How will social statistics in 2000 in the United States compare with social statistics as we know them in 1976? In the attempt to answer this question three factors must be considered: first, in the present state of fields of statistics and trends in respect thereto; second, in the anticipated change in the social economic and political milieus in the approximately one human generation that remains in this century; the third, in the continuing advance of the state of the statistical arts--in the design of conceptual frameworks, in the measurement of social, economic and political phenomena, incompilation and tabulation procedures, and in the invention of new techniques of analysis.

Development of Fields of Statistics. At present in the United States economic statistics in contrast with social statistics are better developed, more systematically integrated and more subjected to policy and programmatic use. Economic statistics are more often analyzed in relation to explicit economic goals than are social statistics in relation to explicit social goals. That is, economic statistics serve as economic indicators more than social statistics serve as social indicators. A statistic is an indicator when it is interpreted as measuring progress or retrogression towards or away from an accepted goal.

The history of statistical developments in the United States reveals that over time new inquiries of census schedules and new current series of data, whether based on administrative records or sample surveys, reflect the transition of the country from a rural, agrarian, folk, to an urban, industrial mass society.[1] The emergent, more complex, interdependent, and vulnerable economic and social orders have required increased government intervention in the exercise of planning and regulatory and evaluative functions. The increase in such functions called for more and more data as a basis for policy formation and administrative action. Since I have documented this assertion elsewhere,[2] I shall use only a single illustration for purposes of clarity here.

In the Census of the United States there was no systematic effort to measure unemployment and the total work force as of a given time period until after the unprecedented level and duration of unemployment of the deep depression of the 1930's. The efforts of the government to alleviate the distress of the unemployed and to cope with other aspects of that depression led to a variety of survey experiments--local, state, and national--which resulted in the abandonment of the "gainful worker" approach and the adoption of the "labor force" or "active population" approach to the measurement of the labor force and its employment status. The adoption of the labor force approach in the 1940 Decennial Census was followed by the monthly series of labor force data based on a sample survey as reported in the Current Population Reports or more specifically in the Monthly Report on the Labor Force.[3]

In similar fashion in the transition from an agrarian to an urban society, emergent problems which required national attention led to new census inquiries, new sample surveys, or new administrative statistics, as new agencies were created to deal with the new problems. Among the statistics which proliferated were data relating to internal migration, housing, income, fertility, education, and place of work.

In respect of economic statistics the development of the National Statistical Accounts, not only in the United States but in the world as a whole, reflects the changing character of the economic order and the growing interest and intervention of government in the operations of the economy. In the United States the Employment Act of 1946, which created the Council of Economic Advisers and required the annual Economic Report, certainly greatly stimulated the development and use for policy and programmatic purposes of economic statistics.

Increasing concern in the U.S. with social problems such as intergroup relations; the "urban crisis;" delinquency, crime, and the administration of criminal justice; welfare; health and medical care; education and recreation has given rise to proposals for a Council of Social Advisers to parallel the Council of Economic Advisers.[4] Should such an agency be established there can be no doubt that social statistics would be subjected to stimulation similar to that of economic statistics for further development and to increased policy and programmatic use.

The increase in the scope and use of statistics has, of course, not been confined to government. Similar developments have taken place in the private sector in business, labor, welfare, civic, and educational organizations. Especially notable, by reason of their impact on public policy, are the public opinion polls such as those conducted by Gallup and Harris.

Anticipated Social Change. Next let me state explicitly my assumptions in respect of changes in the social, economic, and political milieu of the United States during the remainder of this century:

1. The transition from an agrarian, folk, to an urban, mass society will continue.

2. The frictions of this transition will be exacerbated and require increasing government surveillance and intervention. The increasing role of government will probably be more resisted in the United States with its greater addiction to its inherited frontier psychology and old economic religion than has been the case in other advanced countries.

3. Government interest in the realm of the "social" as distinguished from the "economic" will increase to match government interest in the "economic."

4. In consequence, government will adopt more explicit social goals to complement present economic goals.

5. The need to measure progress or retrogression in respect of social goals will resulat in a proliferation of social statistics and in efforts to integrate them.

6. Social statistics will, therefore, tend in use increasingly to become social indicators.

7. Governments in the United States will be engaged in more central (as well as regional) state, and local planning and action to supplement and complement the play of market forces in

dealing with collective, as distinguished from private, problems and needs.

8. By 2000, it is assumed that the priority of social over personal rights will have been so established and that the use of the computer in the preservation of privacy and confidentiality will be so advanced that comprehensive data banks will have been created. These data banks, described further below, will make it possible to collate significant information for individuals, households, and institutions on a local, state, and federal level for statistical purposes, while safeguarding privacy and preserving confidentiality. Regulatory and administrative agencies will have access to the data bank but will not be able to obtain more information about individuals, households, or institutions than they now posses or should possess in light of the future developments. The information collected for administrative and regulatory purposes, however, would be available for statistical purposes.

State of the Statistical Arts. What statistics are available at any point in time depend in part on existent conceptual frameworks and what phenomena can be measured. Gross National Product statistics, for example, did not exist until the necessary conceptual apparatus was developed and the necessary components could be measured. Similarly, although the need for statistics on underemployment has been discussed for many years only scanty data on this subject are available because of conceptual and measurement problems. In efforts to measure levels of living, areas of concern have been mentioned from time to time for which statistics were desirable but little, if any, data exist in respect of them because of difficulties of measurement. Examples of such items are "human freedoms," "health," "security," "opportunity," and "happiness." Also resistant to measurement has been the synthesis or integration of social statistics into a single index comparable to GNP in the realm of economic statistics.[5]

It may be anticipated that as measurement techniques improve new statistics will be developed to include the types of items considered above. Improved measurements may be anticipated on aspects of personality, on attitudes and values, and on other psychological, social psychological, and sociological phenomena which will be of public concern.

The advent of the computer has, of course, greatly and positively affected the scope, timeliness, and quality of data. New more powerful and more efficient generations of computers yet to come will continue to have similar effects.

Finally, new techniques of analysis may be expected to influence statistics of the future. Such innovations as log-linear models for multivariate analysis;[6] the proposed "demographic accounting" procedures;[7] a "health accounts" system;[8] and analytical models of various types will undoubtedly influence future statistical developments.

Central Data Banks by 2000. In view of the above considerations, the most important single development in statistics by 2000 will be the emergence of comprehensive data banks for individuals, households and institutions. By that time in the United States a file will be initiated for each individual beginning with a birth certificate, including a record of every significant event in realms in which government has an interest or program, and closing with a death certificate. Among the files which would be continuously maintained would be a medical file--immunizations, disease episodes, contacts with physicians and paramedical personnel; health facilities used and outcomes. The file would also contain information on marriage, divorce and re-marriage; schools attended with fields of concentration and certificates, diplomas or degrees; employment, unemployment and underemployment; information on internal and international migration; income received, including transfer payments; taxes paid; housing; arrests, indictments, convictions, sentences and institutionalizations. Personal characteristics would, of course, also be included such as age, sex, ethnicity or race, etc.

A similar record would be maintained for households with appropriate entries indicating the person's departures from the household including the new household created with new household formation. Comparable files would also be maintained for institutional households.

I am aware that this prophecy is bound to evoke reactions of consternation and visions of Orwell, if not by 1984, then by 2000. I should state that I am in complete sympathy with the right to privacy and the obligation of governments to maintain the confidentiality of information collected for statistical purposes. But I am convinced of the following:

1. That the existence of discrete record files for diverse purposes, such as vital registration, social security, medicare, internal revenue, census tabulations, voting registration, etc., will prove increasingly costly and inefficient; and impose frustrating constraints in the production of data increasingly required for planning and administrative purposes;

2. It is possibly easier to preserve privacy and maintain the confidentiality of information with the computer, even with the present, let alone future generations of computers, than it was in the pre-computer era;

3. It is conceivable, and it will come to pass, that various safeguards would make it impossible for a Richard Nixon, or a J. Edgar Hoover, let alone lesser men, to have access to any data other than that specifically required and permitted by statute about any person, family or institution that could be used in a way inimicable to their interests. That is, the same data bank could be used by regulatory agencies to obtain the individualized data they need for authorized administrative purposes without their having access to other information in the data bank inserted only for statistical purposes. Simultaneously, the data obtained for regulatory and administrative purposes would be available for statistical use under provisions that would not violate privacy and guarantee confidentiality.

I am convinced that even now, let alone by 2000, the necessary combination of ingenuity and technology can be marshalled to achieve these goals. Such a record system maintained on a

decentralized basis, by city or metropolitan area and by state, with provision for national aggregation would make possible the implementation of Richard Stone's demographic accounting proposal; and such proposals as made by the Committee to Evaluate the National Center of Health Statistics in the U. S. for a national health accounts system. It would make possible statistics with agreed upon periodicities of the stocks and flows of human beings in significant categories and functional units.

Needless to say the establishment of such a record system requires much in the way of advance planning--agreement on standard definitions and practices; and the development of classification systems and taxonomies yet to be devised. The types of problems and considerations involved are discussed at some length by Stone in his OECD volume.[9] Such a data bank could be an economical and efficient way of producing many statistics now being produced in a discrete and overlapping manner which defy integration and synthesis. It also could considerably reduce the items now obtained through censuses and sample surveys and duplicate and overlapping surveys and files; and it would certainly, to a considerable extent, offset the costs of the data bank and the derivation of statistics therefrom.

Finally, it should be observed that similar record systems may be established by 2000 for business and industrial enterprises so designed as to serve government, administrative and regulatory as well as statistical needs. Again it is emphasized that adequate safeguards will have been established to maintain the confidentiality of data for individual enterprises.

Specific Examples of Anticipated New Statistics. The new statistics which will emerge will reflect national priorities as the governments turn successively to deal with various problem areas as they become acute and engage national attention.

Before turning to the purely domestic scene let me first focus on development of statistics generated by international interest in helping developing nations to cope with widespread poverty and to accelerate the advancement of levels of living. By 2000, it is likely that much progress will have been made in the measurement of social and economic development and in the synthesis of an index of development for all nations. Progress in this direction has already been made by the United Nations Research Institute for Social Development.[10] It is conceivable that standardized data collections, a world data bank and common tabulation and analytical procedures will enable each nation to see how it compares with other nations at the same and different levels of development. The United States will be an element in this international statistical system and will have a synthetic index of development as well as component indicators.

Next let us turn to specific domestic developments. On the assumption stated above that the increasingly complex urban, industrial, mass society which will characterize the United States will require increased government surveillance and intervention, major innovations in statistics may be anticipated in areas such as the following: health and medical care; poverty and income maintenance; underemployment as well as employment and unemployment; welfare; minority group status; social mobility; housing; education; crime and the administration of criminal justice and recreation.

Health and Medical Care. By 2000, it may be anticipated that greatly strengthened and broadened health and medical programs will have developed in the U. S. as in other advanced countries, including provisions for comprehensive health insurance or equivalent and improved delivery of medical services. In the development of such national health care system, cooperative health statistical systems will be established linking and integrating present discrete data in the public and private sectors relating to the health industry. The data banks described above would enable a health accounts system to be developed which could relate inputs to outcomes with attention to intermediate processes and flows. Linked and integrated would be such components as vital statistics, health survey data, health manpower and facilities data, medical intervention information, health programmatic data, and medical costs information. Also to be anticipated is the development of data which could make possible the evaluation of medical intervention by relating procedures to outcomes. Needless to say the latter will not be achieved without great controversy. That is, with increasing government interest, surveillance and participation in health and medical care programs, the relation between physician and patient will become a matter of public concern and evaluation of medical practice a consequence.

Poverty and Income Maintenance. By the century's end it will recognized, even in the United States, that poverty has its origins in deficiencies and frictions of the economic and social orders as well as in deficiencies of the family and the individual. Furthermore, it will be recognized that whatever the cause the government will have the obligation to provide an adequate income flow to all persons and families--as far as possible through payments for services performed. It will have long been recognized that "welfare state" is not a pejorative term; that the government must be "the employer of last resort;" and that the problem is not whether the nation is to be or not to be a welfare state, but rather how equitable welfare provisions can be made.

With such orientation it may be anticipated that "poverty" statistics will be greatly strengthened, comparability over time and space much improved, and the many vexing technical problems by reason of changing consumer baskets of goods and inflation reasonably resolved. On the assumption that the United States will have made the elimination of poverty a national goal, poverty statistics will have become poverty indicators and data on income distribution, consumer expenditures, savings and wealth will have become greatly strengthened, routinized and increasingly monitored and used for social as well as economic policy and programmatic purposes.

Labor Force. It has become increasingly evident that the "labor force" or "active population"

approach is not meeting the needs of developing countries for manpower policy and program purposes. One of the reasons for this deficiency lies in the failure of the standard approach to measure underemployment in addition to unemployment. By 2000, it may be anticipated that the measurement of underemployment will have become a standard practice throughout the world in accordance with the recommendation of the Eleventh International Conference of Labor Statisticians in October 1966, held under the aegis of the ILO.[11]

Some indication of the type of data which will become available is afforded by the "labor utilization framework" made operational by the writer[12] for which experimental data in various degrees of completeness are now available for nine nations.[13]

Needless to say, the data bank discussed above will make possible longitudinal as well as cross-section statistics to provide a much better understanding of patterns and changes in labor force participation in relation to other social and economic variables.

Other. Space does not permit even the sketchy information presented above for other statistical areas. In quick summary the following observations are in order:

Housing. Housing statistics will be strengthened and elaborated so that continuous data on stocks and flows will be available; and on quality of housing in relation to occupancy and characteristics of occupants.

Education. Education statistics will be strengthened and planned in significant social and economic context by reason of the central data bank. Measurements will become available on the quality and content of education and on the efficiency and success of schools and educational procedures.

Crime. Data on crime and delinquency and the administration of criminal justice will be improved and integrated and outcomes evaluated. Population surveys and data bank files will provide much more accurate information about the level of criminal and delinquent behavior than obtainable through reports based on police or court records.

Minority Group. The increasing insistence of minority and underprivileged groups, including women, for full equality of opportunity will make much more data on the socio-economic status of such groups available. These statistics will be used as social indicators to measure progress in the elimination of discriminatory practices.

Social Mobility. On the assumptions stated above and the basic assumption that the United States will still be a democracy, it is anticipated that social mobility will be a matter of increasing national concern.[14] In consequence it may be anticipated that periodic statistics will become available on increase or decrease in the social mobility of the population as a whole and on sub-groupings of the population.[14]

Concluding Observations. It is clear that the need for a brief presentation precluded comprehensive coverage of the statistical firmament and permitted only sketchy considerations of the specific areas covered. I have tried to present a framework within which the direction of change in statistics during the rest of this century can be visualized. In the specific area to which I have made some reference it should not be surprising that I have, in the main, concentrated on aspects of social rather than economic statistics. I have done this not only because it is the area with which I am most familiar but, also, because it is the relatively underdeveloped statistical area. The major thrust of my remarks is that with the anticipated changes in the social, economic and political milieus, social statistics will not remain relatively undeveloped.

Without question the most controversial of my prophecies will be that relating to the central data banks. It may be useful to point out that emotional reactions against such a development may be attributable to the fact that the reaction flows from 1976 attitudes and realities not from the attitudes and realities of the year 2000.

It is fitting to close with the thought that should central data banks of the type discussed come to pass, many of the frustrations that face statisticians today will have disappeared; and that this new and greatly enriched source of information will tax the ingenuity of the statistician in producing more and better data in the public interest.

Footnotes

1. Philip M. Hauser, Social Statistics in Use, New York: Russell Sage Foundation 1975, pp. 5-13; also "Social Accounting" in Paul F. Lazarsfeld et. al. The Uses of Sociology, New York: Basic Books, 1967, pp. 839-846.
2. ibid.
3. Bureau of the Census and Bureau of Labor Statistics reports. For example, see U.S. Department of Labor, Bureau of Labor Statistics. Employment and Earnings, vol. 24, No. 7, July 1977. Washington, D.C.:U.S. Government Printing Office, 1977.
4. U.S. Congress, Senate Committee on Government Operations, Subcommittee on Government Research. Hearing on the Full Opportunity and Social Accounting Act (S.843). 90th Congress 1st Sess. Washington, D.C.: Government Printing Office, 1967.
5. Philip M. Hauser, op. cit, Chapter 16.
6. For example, see Leo A. Goodman, "Guided and Unguided Methods for Selection of Models for a Set of T Multidimensional Contingency Tables," Journal of The American Statistical Association, 68 (1973) pp. 165-175. Leo A. Goodman, "A Note on Cohort Analysis Using Multiplicative Models and The Modified Multiple Regression Approach," Unpublished Manu script, Department of Statistics, University of Chicago, 1975.
7. Richard Stone, Demographic Accounting and Model-Building, OECD Education and Development, Technical Reports, Organization for Economic Cooperation and Development.
8. U.S. Department of Health, Education, and Welfare. Public Health Service Health Resources Administration. Health Statistics Today and Tomorrow: A Report of One Committee to Evaluate The National Center for Health Statistics, Vital and Health Statistics - Series 4, No. 15, Washington, D.C.:

Government Printing Office, 1973.
9. Richard Stone, op. cit.
10. D. V. McGranahan, et. al. Contents and Measurement of Socio-economic Development, New York: Praeger, 1972.
11. ILO, International Recommendations on Labor Statistics, Geneva: International Labor Office, 1976.
12. Philip M. Hauser, "The Measurement of Labor Utilization" Malayan Economic Review, Vol. 19 No. 1, April 1974, pp. 1-15. Philip M. Hauser "The Measurement of Labour Utilization-- More Empirical Results." Mimeograph. International Statistical Institute, 1977.
13. Hong Kong, Indonesia, Malaysia, Philippines, Singapore, South Korea, Taiwan, Thailand, United States.
14. For example of study based on such data, see Peter Blau and O. D. Duncan, The American Occupational Structure, New York: John Wiley & Sons, Inc., 1964.

DISCUSSION

Conrad Taeuber, Georgetown University

Those of you who have not known Phil Hauser as long as I have may not find it strange, as I do, to have him express a sense of inferiority about his field. He makes a bow in the direction of economic statistics as being better developed, more systematically integrated, and more subject to policy and programmatic use. One could gain the impression that he overlooks the fact that much criticism is currently being levelled at use of GNP as an indicator of national well being. True, there is no counterpart which clearly points out the gaps in our statistical system, as the GNP has done. There is also less hesitation on the part of public figures to venture into social policy than into economic policy without a firm basis of statistical information. But given the diversity of goals for social policy, it is doubtful that we are ready for one synthetic index which might be thought of as the Comprehensive Social Welfare Index.

It is valid to assume that the growing concentration of our population in clusters which we designate as metropolitan creates a setting in which government and other organizations have a growing role to play. Given the multiplication of human contacts in the urban society, there is little doubt that we will see more governmental intervention and that private organizations of many kinds will play an ever increasing role. All of them will call for more information, for it will become increasingly apparent that the knowledge which an individual gains through his own contacts is insufficient as a basis for action, and that more and more the administrator will recognize a need for an array of firmly established facts.

We have already witnessed changes in the attitudes toward the maintenance of privacy and no doubt there will be further developments in this field. As Hauser points out, the computer has opened up new possibilities and given rise to concerns over the possibilities of abuse, because of its ability to bring together, store, and retrieve vast amounts of information. However, the suggestion of huge data banks available for both administrative and statistical purposes may be going further than we would be prepared to go in the foreseeable future. It is easier to visualize two sets of data collection and storage, with one devoted to statistical activities and the other devoted to individual rights and benefits. The former should have access to the latter but not vice versa. I believe we are going to be willing to pay the price of some duplication in order to maintain this degree of separation.

One possibility which opens new fields for analysis and raises additional fears about the invasion of privacy is that involved in longitudinal studies, i.e., the ability to follow a person, a family, a firm, etc., over a period of time. The cross sectional analyses which are the major sources of information about change cannot do what is possible from an analysis of data that follow a person or a cohort through an appropriate time period. No doubt we will lose some of our sensitivity about providing a basis for such longitudinal analysis, and also lose some of the fear of a data bank which has "everything about everybody." But in saying this I am reminded that in Sweden, with its population register which follows a person from the cradle to the grave, there have recently been widespread vocal concerns over the alleged intrusiveness of the government in seeking information which is considered to be in the private domain. On the other hand, a visitor to that country can hardly refrain from voicing surprise at how little analysis has been done with the rich body of data that is located in the population register.

A great deal of work is required to make even the currently available data sets useful for public policy. No doubt that will be done, for the demands for information are growing rapidly and there is a great need for better ways of extracting information from sets of data.

One may be permitted the hope that by the year 2000 the new generation of policy makers will be less concerned with race and ethnicity than we are now, though concern with minority groups in the society no doubt will continue to be high on our list of national priorities.

If the social changes projected by Professor Hauser actually occur, this will not be because they just happened but rather because actions were taken to adjust to social change or to initiate and promote certain changes. Statisticians also need to play an active role. If the statistical system is to meet its obligations, it is necessary that there be far more attention to underlying assumptions and concepts than has been the case in recent years. There are many situations in which the statistician must help the policy maker define the areas for which information is desired. There is an obligation here which has been inadequately fulfilled in the past and which will require more concerted attention in the future. Too often the matter of reviewing and revising concepts which underlie statistical series has been neglected. There is a need for far more attention to the question whether the definitions used reflect the reality for which measurement is desired. Professor Hauser has called attention to his efforts to provide a more adequate measure of underemployment, a concept which has been largely neglected in the official statistics of this and other countries.

This is only one illustration of the need for rethinking the assumptions that underlie our statistical series. The rate of social change is not likely to be less over the next quarter century than it has been in the last one. The social realities which our statistics are intended to reflect are likely to change at a rapid rate. Unless the statistics are continually adjusted to these changing realities, they do not fulfill their proper function and may actually be misleading.

A new element will enter the official statistical picture between now and the year 2000. That is the provision for a Census of Population every five years instead of every ten, as in the past. The Act providing for that Census carries with it the injunction to take full account of data available from other sources. Much of the discussion preceding the approval of that Census centered on

the expectation that the middecade census would become the focal effort around which a whole program of current statistics would be developed and that the consequence would be a more rational and more effective program of demographic and related data. The realization of the proposed integration of census and current data collection remains to be worked out, but is is clear that there are new possibilities which need careful planning.

The last twenty five years have witnessed significant changes in the social statistics which are becoming available, both as to scope and quantity. There have been marked improvements in the ability to extract information from data collection activities. There is no reason to assume that in this respect we "have gone about as far as we can go."

ANALYSIS OF CENSUS BUREAU NATIONAL HOUSING INVENTORY ESTIMATES

David V. Bateman, U. S. Bureau of the Census

I. Introduction

The purpose of this paper is to analyze the differences in housing unit inventory estimates that have been discovered through a comparison of the Annual Housing National Surveys conducted in the fall of 1973, 1974, 1975, and 1976, with other estimates of the housing inventory. These surveys were conducted by the Census Bureau for the Department of Housing and Urban Development.

Analysis of the inventory estimates indicates that the published Annual Housing Survey's inventory estimate (current independent housing unit estimate) appears to be "too high" when compared with the independently derived 1970 census inventory count adjusted for new construction and units lost from the inventory and an alternative estimator of the inventory derived from the Annual Housing Survey (AHS) itself.

The above differences, and the resulting analysis, will be described in detail along with suggestions for future research and action.

Readers should keep in mind that data presented in this paper are intended for analysis of potential biases in the inventory estimates, and as such may differ from published data.

II. Background

Some background information is needed as a basis for discussing the differences.

A. Purpose of the Survey

The Annual Housing Survey - National Sample estimates described in this paper result from data collected in the fall of 1973, 1974, 1975, and 1976. These surveys were designed to provide a current series of information on the size and composition of the housing inventory, the characteristics of its occupants, the changes in the inventory resulting from new construction and from losses, indicators of housing and neighborhood quality, selected financial characteristics, and the characteristics of recent movers.

B. Annual Housing Survey Inventory Estimation Procedure

The sample design for this survey utilizes the basic Current Population Survey (CPS) design (461 primary sampling unit design) in that virtually the same primary sampling units (PSU's) and enumeration districts (ED's) are used; of course, different households within the ED's are designated for Annual Housing Survey interviewing.

The Annual Housing Survey inventory estimates are derived basically by means of a three-stage ratio estimation procedure. The first and second stages of ratio estimation are only incidental to the problem addressed in this paper. The first-stage adjustment was employed for sample housing units from non-self-representing primary sampling units (NSR PSU's) only and its purpose was to reduce the contribution to the variance arising from the sampling of NSR PSU's. This procedure adjusts for the differences that existed at the time of the 1970 census in the distribution by census region, tenure, and geographic residence of the total housing unit inventory as estimated from the sample NSR PSU's.

The second-stage ratio estimation procedure was only employed for AHS new construction sample units (i.e., sample units built after April 1, 1970). This procedure was designed to adjust the AHS estimates of new construction units to independently derived current estimates for selected categories of new construction units for each of the four regions. These independent estimates were considered to be the best estimates available for the number of new construction units. This adjustment was needed to correct for known deficiencies in the AHS sample with regard to representation of new construction units as well as reducing sampling variation in the sample estimate.

The third-stage ratio estimation procedure is of critical importance for this paper. This ratio estimation procedure was employed for all AHS sample units. The procedure was designed to adjust the AHS sample estimates (i.e., the estimates employing a basic inverse of probability weight, noninterview adjustment factors, and first- and second-stage adjustment factors) to independently derived current housing estimates for four types of vacant housing units, and for 24 residence-tenure-race and sex of head categories for occupied housing units.

The second- and third-stage ratio estimation procedures were repeated in an iterative process in order to bring the AHS estimates into close agreement with both sets of independent estimates (i.e., the independent estimates employed for both the second- and third-stage ratio estimation processes).

C. Current Independent Housing Unit Inventory Estimation Procedure

1. New Construction Inventory Estimates -- The second-stage ratio estimation procedure utilizes independently derived estimates of the new construction housing unit inventory. For conventional new construction housing units, the independent estimate was derived from the Survey of Construction (SOC), a survey of housing unit completions conducted monthly by the Bureau of the Census. For new construction mobile homes, an estimate of mobile home shipments was obtained from The Survey of Housing Starts.[1] This estimate was then adjusted to account for mobile homes shipped and actually occupied as primary residences. (A ratio between mobile homes shipped and put in place for residential purposes was established from the 1970 census for the years 1965-1970.) These independent estimates were used in the 1973, 1974, and 1975 surveys. They were not used in the 1976 survey for most categories, as a coverage improvement program was implemented that theoretically provided complete coverage for new construction units.

2. Total Housing Unit Inventory Estimates -- The third-stage ratio estimation procedure

utilizes independently derived current housing estimates. These estimates are obtained separately for occupied and vacant housing units. The independent estimate of occupied housing units was derived from data based on the Current Population Survey (CPS), a household survey conducted monthly by the Bureau of the Census. The independent estimate of vacant housing units was derived from data based on the Housing Vacancy Survey (HVS), a quarterly vacancy survey conducted as part of the CPS by the Bureau of the Census for the Bureau of Labor Statistics.

a. Occupied housing unit independent estimates obtained from CPS -- These estimates were obtained by the following procedure: (Note reference [1] for a history of changes to this method.)

(1) A weight, which is made up of four components, is associated with every person in the CPS sample. These components are:

(a) A basic weight that is the inverse of the probability of selection.

(b) A weight to reflect an adjustment made for interviews that should have been conducted but were not due to a variety of reasons.

(c) A weight that reflects an adjustment to sample persons located in non-self-representing PSU's only, for the purpose of reducing the contribution to the variance arising from the sampling of these PSU's (first-stage adjustment).

(d) A weight that brings the distribution of the sample persons into closer agreement with independent post-census estimates of the distribution of the population by various age-sex-color categories (second-stage adjustment).

An estimate of occupied housing units is then obtained by summing the principal person's weight for all households. The principal person's weight for a household is defined to be the wife's weight in a husband-wife household or the weight of the head of the household for all other types of households.

(2) The CPS estimates of occupied housing units were obtained for the 35 months preceding the survey date.

(3) A 12-month moving average of the above 35 estimates is then obtained. Twenty-four averages result from this computation.

(4) A least squares regression line is then fitted to these twenty-four 12-month moving averages.

(5) The least squares line is then used to predict what the occupied HU estimates will be for the survey date. The third-stage AHS adjustment, as explained above, is made for 24 different categories of residence-tenure-race and sex of head. The estimate of total occupied housing units, as estimated from the regression line, was allocated to these categories by the following method:

(a) The distribution of occupied HU's from CPS in each of the categories for the four quarters of year of the appropriate survey was obtained.

(b) An average percentage distribution of the occupied HU's was then obtained over the 24 categories for the four quarters.

(c) The percentages obtained in (b) were used to allocate the estimate of total occupied HU's obtained from the regression equation to the 24 categories.

For the most part, this paper will not analyze CPS occupied HU inventory estimates that include the steps described in (3), (4), and (5) above. The purpose of this paper is to study the occupied housing unit estimates coming from CPS, and the regression estimation procedure confounds the analysis.

b. Vacant housing unit independent estimates from HVS -- This independent estimate was obtained by averaging the vacancy estimates from the Housing Vacancy Survey (HVS) for the quarters centered around October of the appropriate survey year. The HVS estimate of total vacants was allocated to the four vacancy status categories by again calculating percentage distributions for the four categories from preceding quarter(s) of HVS for the appropriate survey year.

c. Coverage improvement currently in CPS -- The current independent estimate of occupied HU's is derived from the Current Population Survey, which is supplemented by a sample of units in structures that were missed in the census (E6 Bank), as are most current demographic surveys conducted by the Census Bureau such as AHS, Health Interview Survey, etc. In addition, missed census units and units created since 1970 at addresses listed in the census are picked up by a relisting procedure. The second-stage ratio estimation procedure in CPS, ratios the sample estimates of population to an "independent" estimate of population which has <u>not</u> been adjusted for the undercoverage of persons in the census. Therefore, the coverage improvement in CPS (E6 Bank) has very little effect on the estimate of the number of occupied HU's due to the CPS second-stage ratio adjustment procedure, although it may have an effect on the characteristics. The coverage improvement (E6 Bank) in CPS could affect the housing unit inventory estimate to the extent that these households may be smaller than the average household.

The coverage improvement program conducted as part of the October 1976 Annual Housing Survey attempted to represent the following kinds of units previously not represented in the sample from those areas in permit-issuing jurisdictions where an address sample was taken [4]. These units are presently missing from CPS:

1. Mobile homes put in place as a new address after the 1970 census outside of mobile home parks.

2. Mobile homes put in place after the 1970 census in mobile home parks built after the 1970 census.

3. Mobile homes that were vacant in the 1970 census that have since become occupied.

4. Housing units in structures that have been converted from entirely nonresidential use to residential use since the 1970 census.

5. Housing units that have been physically moved

to a new location since the 1970 census.

6. Housing units for which a permit was issued prior to January 1970, but the units were not completed until after the 1970 census.

7. Mobile homes in parks built before the census but missed by the census.

III. Estimators of the Current Occupied Housing Unit Inventory

Alternative estimators, that are currently available, of the occupied housing unit inventory are the following:

1. The estimator that is used currently to estimate the occupied HU inventory is obtained by creating a smooth monthly time series from CPS. This involves the fitting of a regression line to the monthly CPS estimates. These estimates are presented in column (1) of table A and the methodology is described in detail in section II.C.

2. An estimator of the occupied HU inventory could be obtained by eliminating the regression estimation procedure. This procedure would probably provide poor estimates of change due to monthly variation not attributable to actual inventory changes. Data using this method are presented in column (2) of table A.

3. An estimator of the occupied HU inventory can be obtained from CPS before the second stage of ratio estimation and the regression procedure. These data are presented in column (3) of table A, and include census missed units as well as allocating units picked up in the October 1976 coverage improvement program that was instituted to eliminate known biases in the sampling frame. These data do not have the benefit of a regression procedure and, as such, would probably produce poor estimates of change over time. These estimates are probably somewhat "low" because of two reasons:

(a) Undercoverage in area segments -- Approximately 25 percent of the CPS sample are located in areas of the country where listing procedures are used to obtain a sampling frame. The October 1966 Intensive Coverage Check conducted by the Census Bureau indicated that approximately 1.74 percent of all housing units in area segments were missed by CPS [2]. Thus, approximately 350,000 total housing units could be missed currently in area segments if this result is still reliable. There are indications that this may be an underestimate of the number of HU's missed, as this check was done dependently. For example, the 1970 census evaluation program indicated that over 4 percent of all housing units in rural areas were missed in the census (listing techniques, similar to those used in area segments, were used in rural areas in the census) [3].

(b) The coverage improvement program instituted in the October 1976 AHS survey may have had difficulty in picking up structures that were used for nonresidential purposes at the time of the 1970 census but have since been converted to residential use. The Survey of Components of Change and Residential Finance (SCARF) conducted in 1957-59 and the Components of Inventory Change Survey (CINCH) conducted in the 1960's indicate that we may be missing more units than the 16,000 units picked up in the coverage improvement program.

Table A

	AHS PUBLICATION FIGURE (1) CPS & HVS based inventory est. Includes second-stage adj. and regression estimation (000)	(2) CPS & HVS inventory est. without regression estimation (000)	(3) CPS & HVS inventory without second stage and regression estimation* (000)	(4) AHS "unbiased estimates"* (000)	(5) (2)-(3) (000) CPS	(6) (2)-(4) (000) AHS
October 1973						
Occupied	69,337	69,465	68,173	67,212	1,292	2,253
Vacant	6,632	6,632	6,740	6,608	- 108	24
	75,969	76,097	74,913	73,820	1,184	2,277
October 1974						
Occupied	70,830	71,246	70,060	69,403	1,186	1,843
Vacant	6,771	6,771	6,834	6,730	- 63	41
	77,601	78,017	76,894	76,133	1,123	1,884
October 1975						
Occupied	72,523	72,621	71,653	71,038	968	1,583
Vacant	6,564	6,564	6,627	6,733	-63	- 169
	79,087	79,185	78,280	77,771	905	1,414
October 1976						
Occupied	74,009	73,836	72,423	72,161	1,413	1,675
Vacant	6,528	6,528	6,591	6,828	- 63	- 300
	80,537	80,364	79,014	78,989	1,350	1,375

*Missed HU's have been added back in by means of the coverage improvement program established in the October 1976 Annual Housing Survey, National sample.

4. An estimator of the occupied HU inventory can be obtained from the Annual Housing Survey itself (column (4) in table A). Again census missed units as well as missed housing units obtained from the coverage improvement program are included in these data. These estimates are also probably underestimates for the same reasons cited in 3. above. Theoretically columns (3) and (4) in table A should have the same expected values; however, except for a relatively small problem with 1973 and 1974 AHS estimates, these differences in 1973 and 1974 are unexplainable. In addition to the sources of undercoverage mentioned in 3. above, the AHS in 1973 and 1974 is missing some units that were classified as a loss from the inventory in the regular survey but were found to be legitimate units from the reinterview of lost units.

Again the AHS data do not have the benefit of a regression estimation procedure and, as such, would probably produce poor estimates of change.

5. Another estimator of the inventory that is not included in table A is one constructed from a components of change. Basically, a components of change estimator for a given time period is constructed by adding to the census HU inventory estimate an estimate of the number of new construction units and units added through other means, and subtracting an estimate of the number of units lost from the inventory. At the present time, we are unable to create a good components of change estimator due to the inability of picking up certain kinds of units, namely:

(a) Units that come into the inventory as a result of structures that were used for nonresidential use in the census but have since been converted to residential use and units moved to the present site since 1970.

(b) Units that are going in and out of the inventory over time (flip-flops). For example, you could have a unit that is in the inventory in 1973, a loss in the 1974 survey, and it comes back into the inventory again for the 1976 survey.

(c) Units lost by means of merging units and units added by conversions.

Differences between columns (3) and (4) and column (2) are presented in columns (5) and (6). Note that one cannot interpret these numbers as bias in the CPS estimation procedure, because of the undercoverage problems in area segments, and the inability to obtain certain types of units from converted structures.

IV. Analysis of Current Independent Estimates and "Components" of Inventory Change Estimates from AHS and CPS

Conceptually it seems as though the current independent estimates may be too high for occupied housing units,[2] whereas the independent estimate for vacant housing units may be slightly too low. In addition, a potential problem in estimating the occupied HU inventory could arise from the first-stage adjustment used in CPS; it is unclear at the present time whether this would result in an upward or downward bias.

A. Potential Problem in Occupied HU Inventory

The independent estimate used in the third-stage ratio estimation process for AHS is derived from the CPS estimate of occupied HU's as described previously. The second-stage adjustment procedure in CPS weights up the sample cases without regard to household membership. Thus, the principal person's weight has an additional component that is due to nonprincipal person undercoverage in the survey. In other words, the second-stage adjustment in CPS accounts for both persons in missed HU's as well as persons missed from enumerated HU's. If more persons are missed within enumerated HU's than in missed HU's, the principal person's weight will be biased upward; that is, the CPS might adjust for more undercoverage of principal persons than there actually exists for purposes of estimating the occupied housing inventory.

Model for Potential Bias of CPS Estimates of Occupied Housing Units:

Let the ratio estimate factor used for the second-stage ratio estimation procedure in CPS for a particular age-race-sex cell be represented by

$$f = \frac{P + W}{p' + w'} = \frac{Z}{z'}$$

where:

- Z is an independent demographic estimate of the population for the cell in question
- z' is an estimate of Z derived from the CPS through the first stage of ratio estimation
- P is an independent estimate of principal persons which is unknown
- W is an independent estimate of nonprincipal persons which also is unknown
- p' is the sample estimate of P through the first stage of ratio estimation. This can be obtained from the sample.
- w' is the sample estimate of W through the first stage of ratio estimation. This also can be obtained from the sample.

It is apparent that the second-stage factor that should be applied to the principal person's weight for the purpose of estimating the occupied housing inventory is

$$f' = \frac{P}{p'}$$

1. Note that if $\frac{W}{w'} = \frac{P}{p'} = K$

then $f' = \frac{P + W}{p' + w'} = \frac{Kp' + Kw'}{p' + w'} = K$

and the current estimation procedure is "unbiased" in estimating the occupied housing unit inventory.

2. If $\frac{P}{p'} < \frac{W}{w'}$ which might be the case currently, then $Pw' < p'W$

$$Pw' + p'P < p'W + p'P$$

$$P(w' + p') < p'(W + P)$$

$$\frac{P}{p'} < \frac{W + P}{w' + p'}$$

and the current estimation procedure overestimates

the housing unit inventory. Note table B below for estimates of these factors for the indicated surveys.

Table B

	$\left(\frac{W + P}{w' + p'}\right)$ Average CPS second-stage factors	$\left(\frac{P}{p'}\right)$ * Estimate obtained from coverage improvement program
October 1973	1.03791	1.01968
October 1974	1.02924	1.01229
October 1975	1.02608	1.01252
October 1976	1.03228	1.01311

*Based on October 1976 coverage improvement program in AHS National. Note that the true $\frac{P}{p'}$ probably lies somewhere between these two sets of numbers because of undercoverage in area segments and possible weaknesses in the ability to pick up certain kinds of units such as units in structures converted from nonresidential to residential use.

3. If $\frac{P}{p'} > \frac{W}{w'}$ then it can be shown as in 2. above that $\frac{P}{p'} > \frac{W + P}{w' + p'}$

and the current estimation procedure underestimates the occupied housing unit inventory.

B. Potential Problems in Estimation of Vacant HU Inventory

The independent estimate, of the vacant HU inventory, used in the third-stage ratio adjustment process in AHS is derived from the HVS estimate of vacant HU's. The HVS estimation procedure does not have a second-stage ratio estimation procedure. Therefore, the coverage improvement in CPS (E6) as well as the units picked up in the relisting procedure would appear to adjust for the undercoverage of vacants in the census. One should also note that current surveys have a better coverage rate of vacant units than does the census. Therefore, the independent estimate of vacant HU's would appear to be conceptually correct except for undercoverage due to frame deficiencies [4] explained in section III, and due to undercoverage in area segments.

C. Comparison of April 1970 CPS Estimate of the Occupied HU Inventory and the 1970 Census Count of the Occupied Inventory

Up until the present time, the only validation of the CPS occupied HU inventory estimate has come from the 1970 census inventory count. The inventory estimates for the CPS and the census were relatively close:

1970 census occupied HU inventory count adjusted for undercoverage ---------- 64,338,000

1970 census occupied HU inventory count unadjusted for undercoverage ---------- 63,450,000

April 1970 CPS occupied HU estimate[3] - 62,971,000

This evidence would seem to indicate that there are no problems with the present procedure in estimating the occupied inventory. Certainly there is no evidence here of a tendency to "overestimate" the inventory. Nevertheless, one has to be careful in interpreting these numbers; the CPS procedure could have "overestimated" the occupied inventory for April 1970 and possible "offsetting biases" eliminated a potential bias. Possible "offsetting biases" are:

1. Because of certain kinds of frame deficiencies in the CPS sample, the missed rate of principal persons could have increased over the decade approaching the missed rate of nonprincipal persons.

2. During the sixties and early seventies, an additional step was employed after the least squares method to obtain an occupied HU inventory estimate. The incremental change in the occupied housing unit inventory from the previous census to the most current month represented on the least squares line was calculated using the following formula:

$$I = \frac{Y-H}{N}$$

where:

Y = the number of occupied housing units as estimated from the regression line for the most current month that was used as input to the calculation of the regression line

H = the number of occupied housing units in the 1960 (1970) census

N = the number of months that have elapsed between April 1, 1960 (1970) and the most current month that was used as input to the calculation of the regression line

In order to project the regression line to the current quarter of interest, the value (4.5) times (I) was added to the occupied housing unit inventory estimate that was read off the least squares regression line, for the most current month that was used as input to the calculation of the regression line. The overall effect of this procedure is to dampen the projected inventory estimates.

3. The 1970 census had less undercoverage of occupied housing units than the 1960 census. A conservative estimate is that the 1960 census missed an additional 250,000 occupied housing units over the 1970 census. Therefore, one would expect the inventory estimate built up from the 1960 census to fall short of the 1970 census by this amount if this was the only problem occurring.

D. First-Stage Ratio Adjustment Used in CPS and Its Potential Effect on Estimating the Occupied HU Inventory

The CPS first-stage ratio adjustment procedure may have some effect on the housing inventory estimate. This procedure was employed for sample persons from non-self-representing (NSR) PSU's only. The procedure was designed to reduce the contribution to the variance arising from the sampling of NSR PSU's. This ratio adjustment takes into account the differences that existed at the time of the 1970 census in the distribution by region (four regions), SMSA's (Central Cities, balance urban and balance rural), outside

SMSA's (urban, rural nonfarm and rural farm), and race (white and all other races). The first-stage ratio estimate for each specified category was as follows:

$$\frac{\text{The 1970 census population in the particular category for all NSR strata}}{\text{Estimate of the population in the particular category using 1970 census counts for sample NSR PSU's}}$$

The numerators of the ratios were calculated by obtaining the 1970 census population counts for each of the categories for each NSR stratum and summing these counts in a particular category across the NSR strata in each region. The denominators were calculated by obtaining the 1970 census population counts for each of the categories for each NSR sample PSU, weighting these counts by the inverse of the probability of selecting that PSU and summing these weighted counts in a particular category across the NSR PSU's in each census region. The computed first-stage ratio estimate factor was then applied to the existing weight for each NSR sample person in each first-stage ratio estimation category.

A problem could exist if these sample PSU's are characterized by unusually large or small households. If the PSU's are characterized by large households, the first-stage ratio adjustment process would yield underestimates of the occupied housing unit inventory and in the case of small households it would yield overestimates. It seems intuitively probable that these biases could cancel out over the 220 NSR strata used in CPS. At the present time it is felt that this problem is not a "major one" and thus research as to the extent of error is not designated as a major project although consideration is being given to revising this adjustment for the purpose of obtaining HU inventory estimates.

E. Analysis of "Components of Inventory Change" Estimates for October 1973, 1974, 1975, and 1976

Estimates of change in the total housing inventory are published for two time periods; one period uses the 1973 survey as the base, and the other uses the 1970 census estimate.[4] A key element in these tables are unspecified units. Unspecified units are the difference between

(1) the present year survey estimate and

(2) the base year estimate adjusted for new construction HU's and HU's lost from the inventory.

These units reflect additions to the inventory which are not specifically sampled for the survey, offset by certain losses. Such additions include conversions, changes from nonresidential use, housing units moved to site, units returned to the inventory in a particular survey year, etc., which were definitional losses in a prior year (for example, mobile homes which were vacant in 1973 but became occupied in the particular survey year). Examples of certain losses are mergers and mobile homes occupied in a particular year and vacant in succeeding years, etc.

Table C shows unspecified unit counts for the present published housing unit inventory, and for alternative estimators of the inventory obtained from the Current Population Survey (CPS) and the Annual Housing Survey (AHS). Examination of this table shows that the alternative estimators overall yield slightly smaller counts of unspecified units for comparisons using the 1970 census as a base. However, there is no apparent reduction when the 1973 survey is used as a base.

Table C

UNSPECIFIED UNITS (000's)

Time period	Final CPS based est. (published)	Revised CPS based est.*	AHS unbiased est.*
1970-73	195	-861	-1,954
1970-74	944	237	524
1970-75	1,820	1,013	504
1970-76	1,808	285	260
1973-74	749	1,098	1,430
1973-75	1,527	1,776	2,360
1973-76	1,613	1,146	2,214

*CPS and AHS occupied HU's are calculated by eliminating the second-stage ratio estimation stage and restoring missed housing units obtained in the October 1976 AHS coverage improvement program.

V. Conclusions

Associated with a national sample survey that is conducted to estimate the occupied housing unit inventory are a "target" population and a "survey" population. The target population comprises the total set of occupied housing units in the country and the survey population comprises the entire group of occupied housing units included in, or associated with, the frame and the estimation procedure. One of the purposes of the Annual Housing Survey is to produce an estimate of the number of occupied housing units in the target population. Because of the nature of the second-stage ratio adjustment procedure used in CPS, in which no adjustment is made to the independent estimates of population for undercoverage of persons in the census, the survey population does not account for occupied housing units missed in the census. However, due to a conceptual bias in the CPS estimation procedure, explained in section IV, that overestimates the number of units in the survey population, this procedure may inadvertently provide good estimates of the number of units in the target population. In fact, the present procedure may be producing "better" estimates of the number of units in the target population than presently available alternative estimates described in section III.

Nonetheless, in any survey operation errors of methodology should be corrected when they are discovered. At the present time we are unable to prove conclusively that any of the available estimators are better than the estimator currently being used. However, we believe a better estimator of the occupied housing inventory might be obtained from the CPS by eliminating the second-stage adjustment (in CPS), introducing a coverage improvement program similar to the program introduced in the AHS, improving the coverage of

housing units in area segments, and introducing a more sophisticated regression model to produce better estimates of change.

VI. Possibilities For Future Research and Potential Action

It is generally felt that the present procedure of ratio adjusting the Annual Housing Survey inventory estimates to the independent estimates generated by CPS and HVS is a desirable one if the CPS and HVS estimates can be refined. A refined CPS (monthly estimates) and HVS (quarterly estimates) could provide a smooth series of inventory estimates over time yielding good estimates of change. Since estimates of change are deemed extremely important for the Annual Housing Survey, and since it seems advisable to therefore continue ratio adjusting the AHS to the CPS and HVS inventory estimates, a future revision in the methodology used to estimate the CPS occupied HU inventory may also need a historical revision in the series back to the 1970 census. This revision could include the following research and actions:

1. A re-running of all CPS monthly tapes back to the 1970 census to reconstruct occupied housing unit inventory estimates that are not conceptually biased by the second-stage factors. This would involve the installment of a coverage improvement program in the Current Population Survey. Both month and year built information would have to be collected in order to allocate these units to the proper month CPS would have initially picked them up.

2. The installment of a new modeling technique to predict the occupied inventory for a given month. At the present time a simple regression line of the form $y = a + bt$, where t is a time index and y represents the 12-term moving average (explained in section II.C.2.a.) is used. The effect of the current form of regression is to project forth a linear trend where the objective should be to project a trend cycle. The effect of this adjustment on an inventory that we know has a strong cyclical movement could distort the estimate depending upon where the current inventory is in the business cycle. A new regression technique might allow for curvilinear movements in the occupied inventory estimates.

3. If we shift to a revised CPS estimate of the occupied HU inventory, the only theoretical undercoverage would exist in areas where area sampling procedures are employed. Therefore, coverage procedures could be "firmed" up by testing alternative listing procedures, etc. The present procedure of listing is primarily an "observational" one in that listers are instructed to inquire at the unit only if they are unable to obtain an address from readily visible outside sources. Inquiry at all units would pick up additional units at an additional cost.

Another source of possible undercoverage could be structures in address segments that were classified as nonresidential in the census but have since been converted to residential use. The coverage improvement program instituted in the October 1976 AHS survey picked up relatively few of this type of unit. It is generally felt that the successor method that was used may be ineffective in picking up this kind of unit. Note Montie's and Schwanz's paper for more information on the coverage improvement program [4].

4. Revised first-stage ratio estimate factors for CPS could be calculated for purposes of estimating occupied housing unit inventories. These factors would be based upon 1970 census housing unit distributions. The present procedure is acceptable for measuring population characteristics.

5. Research might be conducted into the possibility of developing a new system of principal persons and nonprincipal persons population controls. If this could be done, undercoverage in area segments and undercoverage due to conversions would not be so critical for inventory counts.

6. The Components of Change Inventory Estimator may be a good source of future inventory estimates if techniques can be devised to estimate certain troublesome components that were elaborated on in section III.

FOOTNOTES

[1] Construction Reports, Housing Starts. C20-76-11.

[2] Morton Boisen, formerly of the Census Bureau, suggested this may be the source of a bias in the estimation procedure.

[3] Based on aged 1960 population controls.

[4] Current Annual Housing Reports. Series H-150-75A.

REFERENCES

[1] Shapiro, Gary M. Internal Census Bureau memorandum addressed to Charles D. Jones, January 14, 1977. Subject: "Modifications to the Procedure for Establishing Quarterly Control Totals of Occupied Housing Units by Tenure for the Quarterly Household Survey and A Description of the Procedure to be Used to Generate Control Totals of Vacant Housing Units."

[2] U.S. Bureau of the Census. The Current Population Survey Reinterview Program, January 1961 through December 1966. Technical Paper No. 19.

[3] Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1970, The Coverage of Housing in the 1970 Census, series PHC(E)-5.

[4] Montie, Irene and Dennis Schwanz. Coverage Improvement in the Annual Housing Survey. 1977 American Statistical Association Proceedings, Social Statistics section.

COVERAGE ISSUES RAISED BY COMPARISONS BETWEEN CPS AND ESTABLISHMENT EMPLOYMENT

Alexander Korns, U.S. Department of Commerce 1/

Labor market analysts have long been puzzled by the fact that the Census Bureau's Current Population Survey (CPS) measure of nonagricultural wage and salary employment declines less in business contractions than does the Labor Department's payroll (or establishment) survey measure. 2/ In this paper, I will show that a major cause of the difference in the cyclical behavior of the two employment measures is the fact that the CPS does not cover about 6 percent of the population.

In chart 1, I have plotted the seasonally adjusted difference (DIFF) between the adjusted payroll and CPS measures of nonagricultural wage and salary employment, monthly for the 20-year period from 1956 to 1976. The two adjusted employment measures were derived as follows:

1. The adjusted payroll measure equals the published measure, minus a small number of employees in agricultural services.

2. The adjusted CPS measure equals the published CPS measure of nonagricultural wage and salary employment, plus 14- and 15-year-old nonagricultural wage and salary workers outside private households, minus those workers age 16 and over who either work in private households or are on leave from their jobs without pay.

DIFF -- Throughout the period between 1956 and 1976, the difference between the two adjusted employment measures averaged about 4 million. This is largely -- but not completely -- attributable to two factors.

1. The payroll measure is conceptually larger than the CPS one, because the payroll survey counts jobs, whereas the CPS counts workers. Thus, the payroll survey counts the second (and subsequent) jobs of multiple jobholders, and this factor contributed roughly 2 million to the average level of DIFF.

2. The CPS measure understates employment, because the independent population control totals understate the population age 14 and over by roughly 4 million -- due to census undercount. This statistical error in the CPS contributed roughly another 2 million to the average level of DIFF.

The focus of my work is not, however, on the average level of DIFF; it is, rather, on the variation in DIFF over time. For purposes of comparison with DIFF, I have plotted the adult male unemployment rate in chart 1, with a dashed line, on an upside-down scale. You can see in the chart that there is a cyclical pattern to DIFF: every time there is a business contraction and the unemployment rate rises, DIFF declines. Conversely, when business recovers and the unemployment rate falls, DIFF generally increases.

This cyclical pattern to DIFF reflects the following situation. In contractions, the adjusted payroll employment measure declines more than the adjusted CPS employment measure. For example, from August 1957 to April 1958, the adjusted payroll measure declined 2.1 million, while the adjusted CPS measure declined only 1.2 million; consequently, DIFF declined 0.9 million (all figures seasonally adjusted). In recoveries and expansions, the adjusted payroll measure generally increases more than the adjusted CPS measure.

In principle, the cyclical pattern to DIFF could be due to any of three causes, or to a combination of them:

1. Conceptual differences in the coverage of the two adjusted employment measures could be responsible.

2. Statistical error in the payroll survey might cause the adjusted payroll measure to exaggerate employment fluctuations.

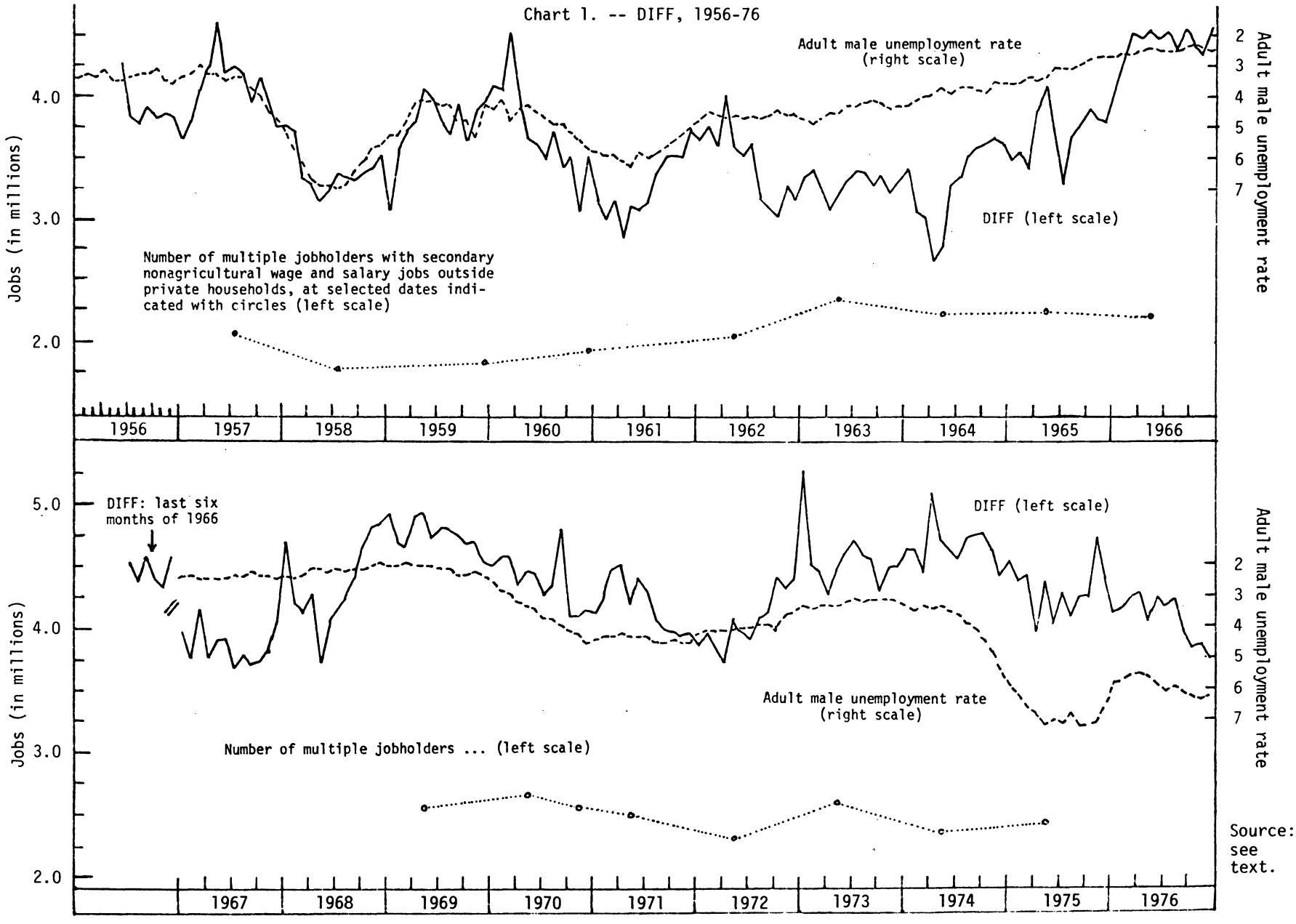
3. Statistical error in the CPS might cause the adjusted CPS measure to dampen employment fluctuations.

In my research, I have examined all of these possible explanations. In this report, I will deal very briefly with the first two, in order to concentrate on the third explanation -- statistical error, and specifically undercoverage, in the CPS -- which is the theme of this panel.

Multiple jobholding and job changing -- First, as previously mentioned, the major conceptual difference between the two adjusted employment measures is that the payroll survey counts jobs, whereas the CPS counts workers. This has two consequences.

1. The monthly CPS classifies multiple jobholders by the characteristics of their primary job -- that is, the job at which the largest number of hours were worked. Consequently, while the adjusted payroll measure counts secondary nonagricultural wage and salary jobs outside private households, the adjusted CPS measure omits such jobs. Intermittent CPS surveys of multiple jobholding indicate that the number of secondary nonagricultural wage and salary jobs outside private households -- plotted with small circles connected by a dotted line in chart 1 -- does not fluctuate sharply with the business cycle. Thus, multiple jobholding does not explain the cyclical pattern to DIFF, although it may contribute to the pattern in a minor way.

2. If a worker leaves a job in a pay period that includes the 12th of the month and starts another job in a pay period that



Source: see text.

includes the 12th of the same month, the payroll survey counts both jobs, whereas the CPS counts one worker. The Social Security Administration's Continuous Work History Sample (CWHS) provides clear evidence that job changing declines during contractions and increases during recoveries and expansions. An illustrative calculation suggests, however, that this factor accounts for cyclical fluctuations in DIFF that are on the order of 50,000-100,000, and are therefore much too small to explain fully the observed cyclical pattern to DIFF.

The payroll series -- Next, I will briefly discuss the hypothesis that the payroll series exaggerates cyclical employment fluctuations. The payroll series is benchmarked, usually once a year, to universe counts of employment based on administrative records. To benchmark employment in the private sector BLS has mainly used ES-202 reports. These are quarterly tax returns submitted by employers to State agencies in compliance with unemployment insurance (UI) laws. On the returns, employers state the number of persons who worked or received pay in the pay period that included the 12th of each month. I will first discuss the reliability of ES-202 data, then the payroll series as a whole.

The principal cause of inaccuracy in the ES-202 data is the attempt by some employers to evade UI taxes by either not filing returns or by concealing some workers. If tax evasion were to increase during business contractions, ES-202 tabulations would exaggerage cyclical employment declines. Realistically, tax evasion is feasible only for very small firms. Cyclical declines in payroll employment have been concentrated almost entirely in goods-producing industries -- manufacturing, construction and mining. If evasion does increase during contractions, the increase would have to be concentrated among small firms in goods-producing industries. But data for the most recent contraction indicate that the increase in evasion among these firms cannot have been very large. From March 1974 to March 1975, goods-producing firms with fewer than 20 workers reported an employment decline of only 5.0 percent, or 129,000 workers, on ES-202 returns; meanwhile, all firms in goods-production reported a decline of 11.4 percent, or 2.56 million workers. 3/

BLS supplements ES-202 reports with other data sources to benchmark employment, and uses data from a panel of 160,000 establishments to interpolate employment for months between benchmarks. There are problems with these procedures: Some of the other benchmark sources may be less reliable than ES-202 reports; and the panel, which is not a probability sample, may introduce bias into the inter-benchmark estimates. However, for the private sector, a comparison of the payroll series with ES-202 tabulations for the same months indicates that the two series have moved in parallel over the course of each business cycle. Therefore, if you accept the ES-202 tabulations as an accurate measure of cyclical changes in employment in firms covered by UI laws, it follows that the payroll series has not exaggerated cyclical employment fluctuations in the private sector. It is implausible that error in the payroll series for government employment has substantially exaggerated cyclical fluctuations in total payroll employment, because government employment continued to grow at all phases of the business cycle.

Error in the CPS

I come now to the central thesis of this report -- that statistical error in the CPS has dampened cyclical employment fluctuations. Much of the analysis will take place in terms of employment ratios. The aggregate employment ratio is the percentage of the civilian noninstitutional population (CNIP) age 16 and over that is employed; similarly, for any sex-race-age group, the employment ratio is the percentage of the CNIP in the group that is employed. 4/

Data from two independent sources underlie the monthly CPS employment estimate.

1. From the most recent decennial census, the Census Bureau extrapolates current population control totals for 96 separate sex-race-age groups.

2. From the monthly sample of 47,000 households, the Census Bureau ascertains employment ratios for each of the 96 sex-race-age groups.

To estimate employment, the Bureau blows the sample employment ratios up to the population control totals. There are two flaws that impair the accuracy of the CPS employment estimate.

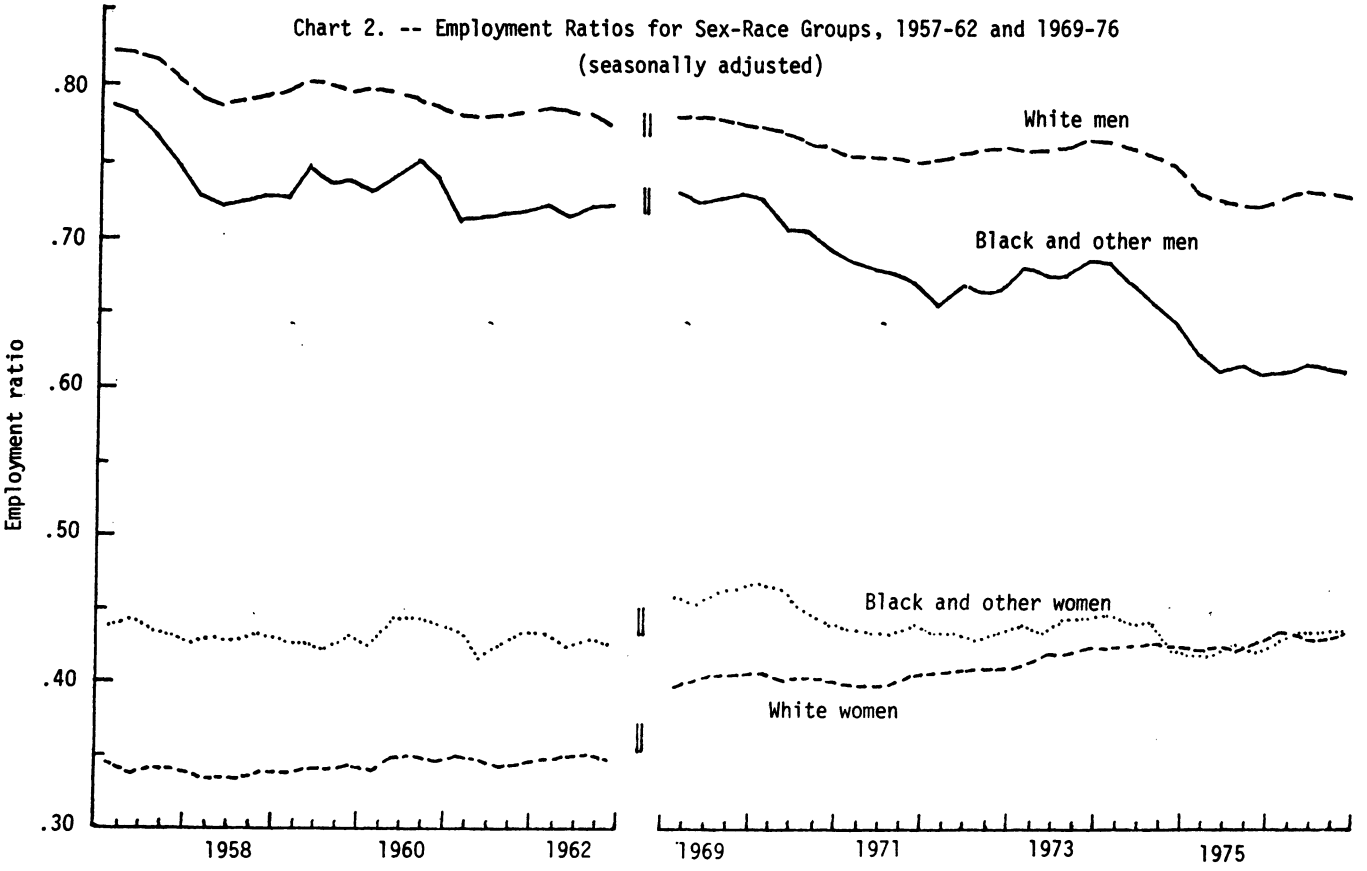
1. The population control totals are too low due to undercount in the decennial census, and the percentage error varies among sex-race-age groups.

2. Sample data for each sex-race-age group are probably biased, because the sample misses some of the persons it is designed to cover.

I will examine the effect of these flaws on cyclical changes in the CPS employment estimate in two steps. In the first step, I will examine the effect of error in the population control totals, on the assumption that the sample data are unbiased. In the second step, I will examine the effect of bias in the CPS sample, on the assumption that the population control totals have been corrected for census undercount.

Control total error -- Jacob Siegel has estimated that the 1970 Census undercounted the population by 2.5 percent. The undercount of the working-age population, 18 to 64, is of particular interest, because this group accounts for almost all employment. The overall undercount rate for this age group in 1970 was 2.8 percent; it was 4.1 percent for men and only 1.5 percent for women. Within each sex group, the rate was about

Chart 2. -- Employment Ratios for Sex-Race Groups, 1957-62 and 1969-76
(seasonally adjusted)



Source: Bureau of Labor Statistics, quarterly averages.

4 times as high for black and other races as for whites. 5/

I define the "control-total-corrected" employment estimate as the estimate that the Census Bureau would have made if it had blown CPS sample data up to control totals corrected for census undercount. In business contractions, control-total-corrected employment would have declined more than published CPS employment, for two reasons.

First, the population base would have been larger. Second, and analytically more interesting, the aggregate employment ratio would have declined more, because those sex-race groups with the largest census undercount rates have experienced the largest cyclical declines in the employment ratio (chart 2). In each contraction, the employment ratio declined far more for men than for women, and within each sex group, it declined far more for black and other races than for whites.

In business recoveries, control-total-corrected employment would have increased more than published CPS employment, but for only one reason -- the population base would have been larger. The aggregate employment ratio would not have increased more, because the sex-race groups with the largest census undercount rates did not (after 1959) experience above-average increases in the employment ratio in recoveries (chart 2).

Undercoverage -- Now I will assume that the Census Bureau has corrected the control totals for census undercount, and I will examine the effect of bias in the CPS sample on cyclical changes in the CPS control-total-corrected employment estimate.

The CPS sample is designed to include about 1 housing unit for every 1,400 in the country. At units designated for the sample, interviewers inquire about the employment activities of all household members age 14 and over, except Armed Forces members. To estimate the population actually covered by the sample, the Census Bureau multiplies the population in each sample household by the inverse of its probability of selection and adds the products. Subtracting the covered population from the best estimate of the population -- i.e., the population corrected for census undercount -- you have the uncovered population. Thus, in 1975, the covered population was 154.1 million and the uncovered population was 9.7 million (table 1, lines 4 and 5).

The undercoverage rate equals the uncovered population as a percentage of corrected CNIP. In 1975, the average undercoverage rate was 5.9 percent. The rate has always been much higher for men than for women, and for black and other races than for whites (chart 3). There are two groups in the uncovered population:

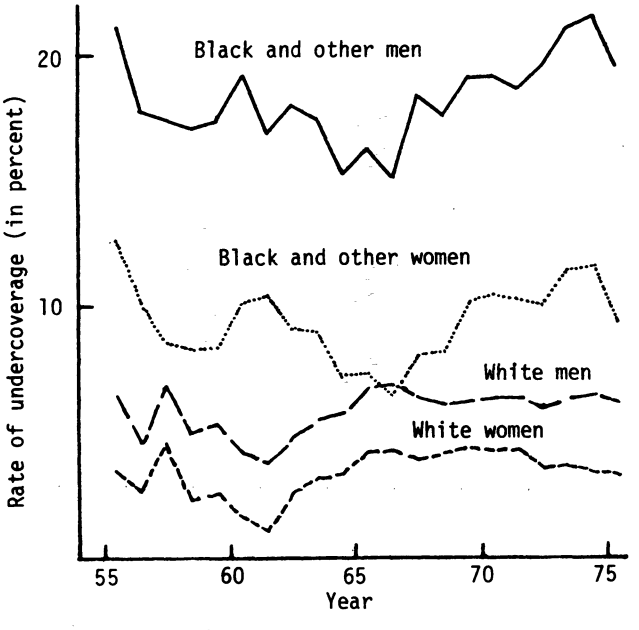
1. The population in housing units missed by the CPS has probably been the minority group at most times. David Bateman will later discuss how the CPS misses housing units. On the basis mainly of information

Table 1.--CPS Undercoverage, 1975 Annual Average (in thousands of persons)

Item	Total	Men			Women		
		Total	White	Black and other	Total	White	Black and other
1. Uncorrected (census-level) civilian noninstitutional population, age 14 and over	159.71	75.70	67.03	8.67	84.02	73.62	10.40
2. Plus: Adjustment for census undercount	4.09	2.73	1.82	.91	1.36	.95	.41
3. Equals: Corrected civilian noninstitutional population	163.80	78.43	68.85	9.58	85.37	74.56	10.81
4. Minus: Population covered by CPS	154.13	72.24	64.58	7.66	81.89	72.09	9.80
5. Equals: Uncovered population	9.67	6.18	4.27	1.91	3.48	2.47	1.01
6. Minus: Population in uncovered housing units	2.91	1.39	1.21	.18	1.52	1.31	.21
7. Equals: Residual uncovered population	6.76	4.79	3.06	1.73	1.97	1.16	.81
Notes (in percent):							
8. Rate of undercoverage -- (5) ÷ (3)	5.90	7.89	6.20	19.98	4.08	3.31	9.38
9. Population in uncovered housing units as a percent of corrected CNIP -- (6) ÷ (3)	1.78	1.78	1.76	1.93	1.78	1.76	1.93
10. Residual uncovered population as a percent of corrected CNIP -- (7) ÷ (3)	4.12	6.11	4.44	18.05	2.30	1.56	7.45
11. Adjustment for census undercount as a percent of corrected CNIP -- (2) ÷ (3)	2.55	3.48	2.71	10.52	1.59	1.29	3.94

Source: Census Bureau.

Chart 3. -- CPS Undercoverage of the Corrected Civilian Noninstitutional Population Age 14 and Over, by Sex and Race, 1956-75



Source: Census Bureau.

he supplied to me, I estimated that in 1975 the CPS missed 2.9 million persons age 14 and over in uncovered housing units, or 1.8 percent of the CNIP (table 1, lines 6 and 9). It is likely that men and women were missed at roughly equal rates in uncovered housing units.

2. The remaining group, what I call the "residual uncovered population," has probably been the majority group at most times. This group consists mainly of residents of covered housing units whom respondents fail to report, for various reasons. The size of this group can only be estimated residually. There were about 6.8 million persons on average in this group in 1975, or 4.1 percent of the CNIP (lines 7 and 10). The miss rate for men (6.1 percent) greatly exceeded that for women (2.3 percent), in this group.

Consequences of undercoverage -- Later, I will present evidence that -- within the uncovered population -- the residual group consists of persons who are poorer on average than their covered counterparts of the same sex, race, and age. For the moment, if you will allow me the assumption that this is the case, I will show that persons in the residual group experience larger cyclical fluctuations in their employment ratios than covered persons of the same sex, race, and age.

Labor economists have long contended that poor

persons suffer disproportionate employment losses during business contractions, and enjoy disproportionate employment gains when the labor market is tight. They believe this happens for two reasons.

1. Relatively few poor persons are in white collar occupations, which experience much smaller cyclical employment fluctuations than do other nonfarm occupations.

2. Employers are said to rank potential employees in a "labor queue." Those persons who lack characteristics that are desirable to employers -- high skill, high educational attainment, and steady work records -- stand at the end of the queue. Poor people lack skills, have low educational attainment, and checkered work records, and therefore stand toward the end of the queue. In contractions, employers are said to lay off disproportionate numbers of workers at the end of the queue; in tight labor markets, when workers toward the front of the queue are not available, employers are said to hire disproportionate numbers of workers at the end of the queue. To some extent, seniority rules reinforce this pattern.

Evidence that provides partial support for this picture of labor market behavior is contained in chart 4. Poverty and low educational attainment are known to be correlated. If poor persons do indeed experience disproportionately sharp cyclical employment fluctuations, I would expect persons with low educational attainment to experience disproportionately sharp employment fluctuations. Men with less than 12 years' schooling have in fact suffered much larger declines in their employment ratio in contractions than men with high school diplomas (chart 4). However, they have not enjoyed disproportionate gains in recoveries and expansions in the period 1964-75. For women -- who constitute a minority of residual missed persons and who are not represented in chart 4 -- the cyclical differentials are similar to those for men, but less pronounced.

For business contractions, the available evidence thus supports the hypothesis that residual missed persons suffered larger declines in their employment ratio than did covered persons of the same sex, race, and age. Consequently, the absence of these persons from the CPS sample dampened the decline in the control-total-corrected CPS employment estimate.

For recoveries and expansions, my findings are less clear-cut. The absence of residual missed persons from the CPS sample may have dampened the increase in the control-total-corrected CPS employment estimate, but this is not supported by the evidence in chart 4.

Illustrative calculation -- In sum, the CPS understates the employment decline in contractions for two reasons: first, because the population control totals are in error, due to census undercount; and second, because the CPS sample is biased, due to undercoverage. Let us return now to the problem I began with -- the cyclical pattern to DIFF. I will present an illustrative calculation of the impact of CPS error on DIFF in the 1974-75 business contraction. From the first 9 months of 1974 to the first 9 months of 1975, the adjusted payroll measure of nonagricultural wage and salary employment declined 1.67 million, while the adjusted CPS measure declined only 1.18 million (both figures are seasonally adjusted). The independently seasonally adjusted DIFF declined 425,000, and CPS error contributed to the decline of DIFF in two ways.

1. Control total error dampened the decline in adjusted CPS nonagricultural wage and salary employment by 105,000.

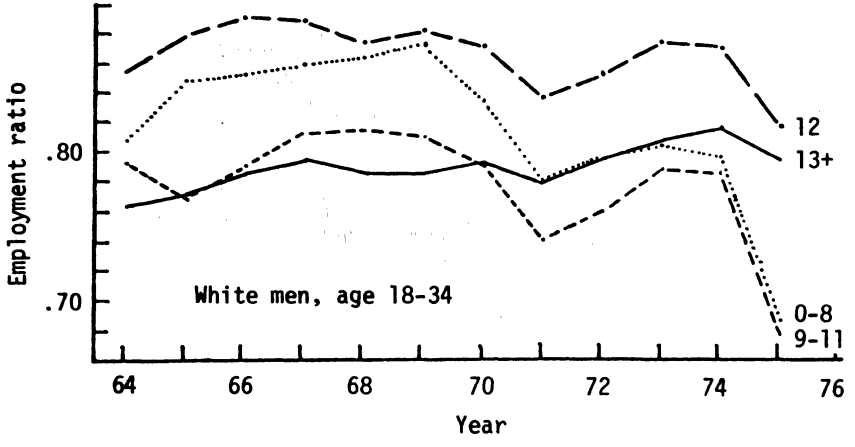
2. CPS undercoverage dampened the decline in control-total-corrected adjusted CPS nonagricultural wage and salary employment by a further amount. I will assume for the sake of argument that: Residual uncovered persons experienced declines in their employment ratios that were twice as large as the declines for covered persons of the same sex, race, and age; and persons in missed housing units experienced the same employment ratio declines as did covered persons of the same sex, race, and age. It follows that undercoverage dampened the decline in the control-total-corrected adjusted CPS measure of nonagricultural wage and salary employment by about 186,000. 6/

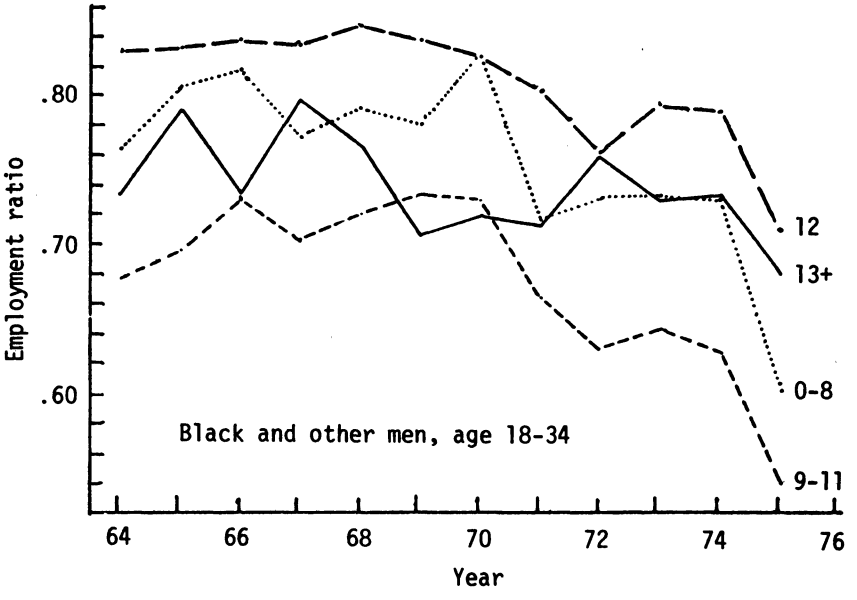
If my assumptions are correct, the two errors dampened the decline in the adjusted CPS measure of nonagricultural wage and salary employment by 291,000. This dampening would explain 68 percent of the decline in DIFF in the 1974-75 contraction.

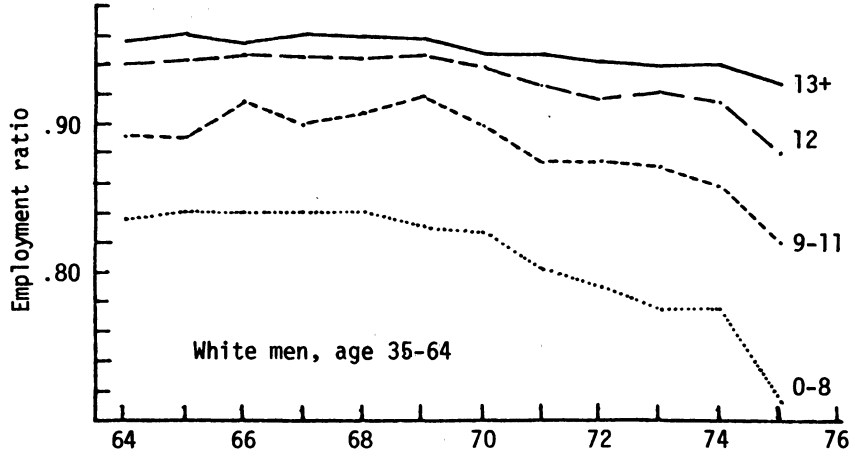
Illegal immigration -- An issue that complicates the analysis of control total error and undercoverage is the effect of illegal immigration on the accuracy of the CPS employment series. The population control totals corrected for census undercount ignore illegal immigrants, because Siegel's estimates of the 1970 undercount and the Census Bureau's estimates of the month-to-month change in the population do not take account of illegal immigration, a subject on which reliable data are altogether lacking. Consequently, the CPS control-total-corrected employment series -- and, ipso facto, the published series -- do not reflect changes in illegal alien employment. 7/ The payroll series, however, appears to count most of the illegal aliens who work in nonagricultural wage and salary jobs outside private households. 8/ DIFF is therefore sensitive to changes in the employment of illegal aliens in nonagricultural wage and salary jobs outside private households; however, DIFF is a somewhat ambiguous indicator of such changes, because it is also sensitive to other factors.

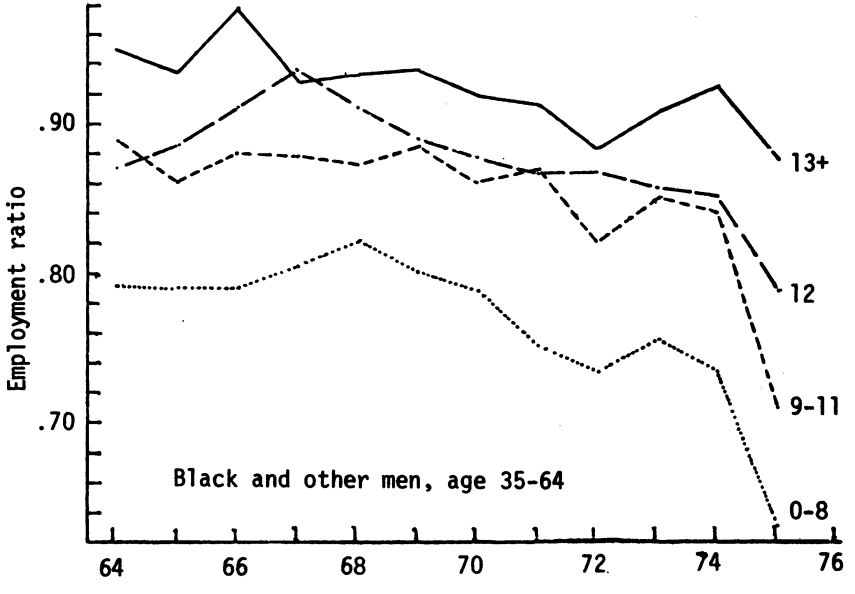
If allowance is made for a break in DIFF in January 1967, 9/ DIFF increased the record amount of 2.2 million from 1964 to 1969 (chart 1). The

Chart 4. -- Standardized Employment Ratio, by Educational Attainment in Years, 1964-75









Note -- Based on unpublished BLS tabulations from the March CPS. Original data for six age groups (18-19, 20-24, 25-34, 35-44, 45-54, and 55-64) were combined. To standardize for secular shifts in the age distribution within each educational group, the average March population for the 12-year period was used as a fixed weight for each sex-race-age-education group.

increase may reflect a sharp rise in illegal alien employment during the long business expansion of 1964-69. I have not been able to identify other factors that could account for the increase in DIFF in 1964-69.

There has been no sustained increase in DIFF since 1970, suggesting that the employment of illegal aliens in nonagricultural wage and salary jobs outside private households may not have increased substantially since 1970, and casting doubt on the widespread impression that illegal alien employment has grown rapidly since 1970. Of course, offsetting factors that tended to reduce DIFF could have masked growth in illegal alien employment, but I have not been able to identify any such factors.

Characteristics of Residual Uncovered Persons

Earlier, I asked you to allow me the assumption that residual uncovered persons are poorer on average than their covered counterparts of the same sex, race, and age. My argument that undercoverage is, in part, responsible for the cyclical behavior of DIFF hinged on this assumption; and it therefore remains for me to present evidence in support of the assumption.

Each month, the Census Bureau provides interviewers with lists of about 55,000 housing units designated for the CPS. There are three ways that interviewers miss persons while canvassing the designated housing units.

1. Interviewers find that an average of 7,500 housing units is vacant or otherwise ineligible for interview each month. Some of the units so classified are actually occupied; the residents of such "false vacancies" are missed.

2. At respondent households, interviewers ask a responsible household member to name all persons "who are living or staying here," including persons who are temporarily absent. Any persons whom the respondent omits from the roster of residents are missed by the CPS.

3. Persons with no usual residence are, of course, automatically missed.

There is evidence that persons missed in each of the three ways are more likely to be poor than covered persons of the same sex, race, and age. For brevity, I will discuss only the characteristics of persons omitted from household rosters. This is by far the largest group among residual uncovered persons.

Incomplete rosters -- Analysts of census undercount and ethnographers have identified two reasons why respondents give incomplete rosters to census enumerators and to CPS interviewers -- concealment and oversight. Both reasons apply with more force to men than to women, and to poor persons than to nonpoor persons.

1. Concealment -- Some respondents fear that information given to the Census Bureau will be used against them, and feel safer in withholding the names of some residents. Men are more likely to be concealed than women for two reasons: first, many of the motives for concealment apply particularly to men; and second, the majority of respondents are women. The motives for concealment are highly correlated with poverty.

Recipients of public assistance have, or may think they have, an incentive to conceal wage-earning or other income-receiving residents. Women receiving Aid to Families with Dependent Children (AFDC) have an incentive to conceal the natural father or adopting stepfather of their children, and may feel safer not reporting a husband or boyfriend even in cases where it would not affect AFDC eligibility. In many States, AFDC recipients also have an incentive to conceal nonearning residents not eligible for AFDC, because welfare officials prorate rent and utilities among all residents in computing AFDC grants.

In an ethnographic study of 35 Puerto Rican households in a poor New York neighborhood, Alan Harwood found the households had failed to report 15 of 52 resident men to a 1967 survey. Whereas the survey indicated that 67 percent of the households were female-headed, Harwood found that only 38 percent were actually female-headed. Fear of losing public assistance was the main motive for concealing the presence of male residents. 10/

Housing regulations create additional incentives for respondents to conceal residents. Fear of police or private retribution is another motive for concealment. Illegal immigrants, persons engaged in illegal activities, and persons wanted by the police have strong incentives to hide. 11/

2. Oversight -- Some respondents apparently "overlook" persons loosely attached to their household when asked to provide a roster of residents. The oversight may be entirely unintentional, or it may be linked to subjective motives such as a wish to protect privacy in regard to ambiguous matters, or a wish to keep interviews short. Ethnographic evidence for blacks indicates that loose attachment to households is more prevalent in poor neighborhoods than in more affluent neighborhoods, and that is more characteristic for men than for women. 12/

Driver's license study -- A Census Bureau study in connection with the 1970 Census provides insight into the circumstances under which respondents in poor neighborhoods omit men from rosters they provide to census enumerators, and, presumably, to CPS interviewers as well. From the rolls of the District of Columbia's Department of Motor Vehicles, the Bureau took a sample of 710 men, age 20-29, mostly black, with addresses in poor neighborhoods and with newly issued or renewed driver's licenses. In attempting to match the names with persons reported to the 1970 Census, the Bureau found that 23.5 percent of the men had been missed or probably missed by the census. There were two groups of missed men.

1. Twelve percent were misses that were confirmed by a resident at the man's address in reinterviews. Of these, 9.0 percent were in housing units that were enumerated and classified as occupied in the census. The investigators were generally unable to obtain clear explanations of why the men had not been reported to the census. Oversight may have been a major reason for this type of miss.

2. The other 11.5 percent were misses or probable misses that residents would not confirm in reinterviews, although the men had received their licenses by mail, and the investigators were frequently able to obtain corroborating evidence from the Post Office or Internal Revenue Service that the men received other mail at the address. Residents said they did not know the men, or said the men lived at other addresses that could not be confirmed in interviews at these addresses, or gave replies that appeared evasive or confused to the investigators. Deliberate concealment appears to have been a major reason for this type of miss. 13/

Sex ratios -- Thus far I have presented impressionistic evidence in support of my assumption that persons omitted from household rosters are poorer than their covered counterparts of the same sex, race, and age.

Table 2. -- Sex Ratios for Persons Age 18 to 64 by Residence in and Outside Metropolitan Poverty Areas, 1975 Annual Average

Race and area	Population age 18 to 64 (millions)	Sex ratio: men per 100 women		Note: percent with 1975 income below poverty level
		CPS sample	Estimated "true"	
	(1)	(2)	(3)	(4)
White.....	107.1	91.8	95.8	9.7
Metropolitan poverty...	4.7	89.1		25.0
Metropolitan nonpoverty	68.2	91.4		7.0
Nonmetropolitan.....	34.2	93.0		12.6
Black and other.....	14.6	75.0	89.2	29.3
Metropolitan poverty...	4.4	68.3		37.4
Metropolitan nonpoverty	7.0	78.3		17.6
Nonmetropolitan.....	3.1	77.8		41.6

Sources and notes:

1. BLS. Refers to census-level civilian noninstitutional population.
2. Estimated sex ratios in the population covered by the CPS, before blow-up of sample data to census-level population control totals. Underlying data from BLS and Census Bureau.
3. Census Bureau. Based on population corrected for census undercount.
4. Census Bureau. Refers to census-level civilian noninstitutional population, all ages, plus Armed Forces members living off base in the United States.

My assumption is also supported by a systematic comparison of sex ratios in the population covered by the CPS in and outside metropolitan poverty areas. The sex ratio is the ratio of men per 100 women. Metropolitan poverty areas are census tracts in which 20 percent or more of the population reported 1969 incomes below the poverty level in the 1970 Census.

Sex ratios in the population age 18-64 that was covered by CPS in 1975 are shown in column 2 of table 2 by race, for metropolitan poverty and nonpoverty areas. You can see that, in metropolitan areas:

1. For whites, the CPS found 2.3 fewer men per 100 women in poverty areas than in nonpoverty areas.

2. For black and other races, the CPS found 10.0 fewer men per 100 women in poverty areas than in nonpoverty areas.

There are two possible explanations for these differences.

1. They may reflect greater CPS undercoverage of men in poverty areas than in nonpoverty areas, due to incomplete rosters.

2. They may reflect lower true sex ratios in poverty areas than in nonpoverty areas.

Although data are lacking with which to settle the issue, the former explanation is more plausible. In defense of the latter explanation, it is sometimes argued that low sex ratios in poverty areas reflect a situation in which men have left their wives and children in poverty areas and gone to live elsewhere. This view is not persuasive, for two reasons.

1. It ignores the findings of ethnographers that many of the households that the CPS counts as female-headed are actually male-headed.

2. It begs the question of where the departed husbands and fathers went to live. Since ethnographers have found that the inability of men to earn steady incomes is a major cause of marital instability among poor persons, it would be surprising if the departed men were to resettle en masse in the more affluent sections of metropolitan areas.

FOOTNOTES

1/ The views expressed in this paper are not those of any Governmental agency. In writing this paper, I have benefited from the generous editorial assistance of Edward Steinberg, and from discussions with David Hirschberg and Fritz Scheuren. Don King and Tom Kraseman first got me interested in the topic. I have received data and other assistance from Paul Armknecht, Carol Utter, and Alan Harwood; and

from Charles Jones, Irv Schreiner, Gary Shapiro, Jacob Siegel, Alfred Tella, Murray Weitzman, and many other persons at the Census Bureau. I would like to thank Patti Trujillo for her charts, Fred von Batchelder for clerical assistance, and Thelma Pearson and Atherine Payne for typing assistance.

2/ See, for example, President's Committee to Appraise Employment and Unemployment Statistics, Measuring Employment and Unemployment, 1962, p. 113.

3/ Employment and Wages, first quarters of 1974 and 1975.

4/ This ratio is higher than the ratio published by the Bureau of Labor Statistics, because the former is based on the civilian noninstitutional population, the latter on the total noninstitutional population (including Armed Forces).

5/ Census Bureau, Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis, PHC(E)-4, 1974.

6/ There are three steps to the illustrative estimate. First, residual uncovered persons accounted for 70 percent of the uncovered population in 1975 (table 1). I therefore assume that the employment ratios of uncovered persons declined 70 percent more than the employment ratios of covered persons of the same sex, race, and age. Second, there were an average of 10.1 million persons age 14 and over in the uncovered population in the two 9-month periods under consideration. If their employment experience had been the same as that of covered persons of the same sex, race, and age, their adjusted nonagricultural wage and salary employment would have declined 267,000. Third, under my assumption, the adjusted nonagricultural wage and salary employment of uncovered persons declined 70 percent more than 267,000, or an additional 186,000.

7/ The population control totals corrected for census undercount ignore most emigration as well as illegal immigration. Therefore, the controls implicitly allow for illegal immigration equal to uncounted emigration. There is some evidence that uncounted emigration during the decade 1960-70 was about 100,000 per year. Robert Warren and Jennifer Peck, "Emigration from the United States: 1960 to 1970," paper presented at the annual meetings of the Population Association of America, 1975; and Ada Finifter, "Emigration from the United States -- An Exploratory Analysis," paper prepared for the Conference on Public Support for the Political System at the University of Wisconsin-Madison, August 13-17, 1973.

8/ There is evidence that employers pay Social Security taxes for about 80 percent of their illegal alien nonagricultural wage and salary employees. Employers who pay Social Security taxes probably pay UI taxes for the same workers when they are covered by UI laws. Consequently, the ES-202 tabulations probably include most illegal alien nonagricultural wage and salary workers outside private households. David S. North and Marion F. Houstoun, The Characteristics and Role of Illegal Aliens in the U.S. Labor Market: An Exploratory Study, report to the Department of Labor, March 1976, p. 142.

9/ In January 1967, BLS reclassified about 750,000 nonagricultural workers from self-employment to wage and salary employment, thus reducing DIFF by the same amount. Note the break in DIFF in chart 1.

10/ Harwood and his associates observed the households for 12-14 months in 1968-69, and reconstructed their rosters as of the date of a 1967 survey conducted by a neighborhood health center. Alan Harwood, "Participant Observation and Census Data in Urban Research," paper delivered at the annual meeting of the American Anthropological Association, 1970; and personal communication to the author.

11/ Leon Pritzker and N. D. Rothwell, "Procedural Difficulties in Taking Past Censuses in Predominantly Negro, Puerto Rican, and Mexican Areas," in Social Statistics and the City, David M. Heer, editor, Report of a Conference held in Washington, D.C., June 22-23, 1967, Joint Center for Urban Studies of the Massachusetts Institute of Technology and Harvard University, 1968, pp. 72-73.

12/ Carol B. Stack, All Our Kin: Strategies for Survival in a Black Community, New York, 1975; and Elliot Liebow, Tally's Corner, Boston, 1967.

13/ Census Bureau, "1970 Census: Preliminary Evaluation Results Memorandum No. 21," prepared by Ralph Novoa, October 1971.

THE IMPACT ON PERSONAL AND FAMILY INCOME OF ADJUSTING THE CURRENT POPULATION SURVEY FOR UNDERCOVERAGE

Robert Yuskavage and David Hirschberg, Bureau of Economic Analysis
Frederick J. Scheuren, Social Security Administration

This paper presents the results of adjusting the Current Population Survey (CPS) for undercoverage, with attention focused on the impact of alternative adjustment procedures on the distribution of personal and family income. In addition, the impact on selected population characteristics and labor force estimates is reviewed.

The data base to which the coverage adjustments were made is the March 1973 CPS. This particular survey was selected because a special matching study conducted jointly by the Bureau of the Census and the Social Security Administration brought together information from the Current Population Survey, the Internal Revenue Service's personal income tax returns and Social Security's wage and benefit systems. 1/ As part of the reconciliation of the differences among these various sources, it was necessary to "correct" for the understatement of the population in the CPS.

Organizationally, the material has been divided into five sections. We begin in section 1 with some background on the nature of the March 1973 CPS' undercoverage and the alternative methods employed to deal with it. The next three sections examine the differential impact of the adjustments on income (sections 2 and 3) and other selected characteristics (section 4). Section 5 provides a few concluding remarks.

1. CPS UNDERCOVERAGE ERRORS AND ALTERNATIVE ADJUSTMENTS CONSIDERED

The papers by Bateman [3] and Korns [4] have already provided a detailed discussion of the nature and magnitude of Current Population Survey and Annual Housing Survey coverage errors. Some further points still need to be made, however, especially with regard to the March 1973 CPS. After sufficient background has been set, we will then describe the alternative coverage adjustments considered.

1.1 Types of CPS undercoverage.--As we have just seen [3, 4], undercoverage errors in a survey or census may be classified into omissions of two types. People can be missed if they live in households that are missed or they can be in an enumerated household but for one reason or another not be counted as members.

In the paper by Korns, estimates from the 1975 CPS were presented to show that for adults, more than two-thirds of the undercoverage was the result of persons missed in enumerated households. This pattern, which seems to have been typical since at least 1975, was not present for the March 1973 CPS due to a number of special circumstances. Instead, in 1973, we estimate that each of the two kinds of error accounted for about half of the undercoverage 2/ --roughly the same ratio that was observed in the 1970 Census [7].

The undercoverage of households in the March 1973 CPS arises mainly from the following sources [8-9]:

(a) Deficiencies inherited from the 1970 Census insofar as the CPS relied on Census addresses as one of the sampling frames (68 percent of the total CPS universe consists of address-list enumeration districts (ED's) drawn from the 1970 Census).

(b) Incomplete listings in area segments; the failure to include established mobile homes in address ED's or to systematically include new mobile homes.

(c) The omission of addresses converted from nonresidential to residential or homes moved to a site which was not a residential address in the 1970 Census; failures to include housing units completed after the Census for which permits were issued before January 1, 1970.

(d) Failures to include addresses imputed in the Census or ones inadequately described in the address registers. (The "Cen-Sup" or E6 portion of the CPS now corrects this inadequacy [3]; however, that sample was not put in place until shortly after March 1973.)

Altogether, the effect of these omissions was to understate the total number of occupied housing units by approximately 3.3 percent.3/ With the addition of the Cen-Sup sample, CPS household undercoverage has been reduced to a rate now less than 3 percent [9].

The sources of within household undercoverage in March 1973 can only be speculated about. Siegel [5] suggests causes of within household misses in the 1970 Census which may be applicable:

(a) Deliberate concealment, carelessness, confusion, or apathy on the part of respondents.

(b) A failure to adequately allow for persons who do not fit into any household according to the conventional rules of residence.

There are some differences between the CPS and the Census that could give rise to other possible sources of error which, when taken together, probably lead to additional within household undercoverage. These are:

(a) Differences in enumeration (i.e., the replacement of self-enumeration with enumeration by highly trained, generally experienced CPS interviewers--possibly leading to some improvement in coverage).

(b) Differences in counting rules. (In particular, college students living in dormitories are counted at their college in the Census. They are supposed to be counted at their parent's home [permanent address] in the CPS. We suspect this change leads to some loss of coverage.)

(c) Longitudinal nature of CPS. (Coverage rates typically decline over the life of a CPS rotation panel. This may be partly due to new household members not being completely accounted for on household rosters in succeeding interviews.)

March 1973 was characterized by exceptionally good overall coverage, better than that for any other March survey in the period 1970-1977. This is true even though the survey's household coverage, as we saw above, was not as good as that for the March CPS surveys since then. One of the main reasons for this apparent paradox may be that in the 1973 survey, because of updates in the sample design, the mix of rotation panels was not the same as that which typically occurs in every other month. The average number of previous interviews each household had received was considerably less than normal in March 1973.

1.2 Alternative coverage adjustments.--A coverage adjustment has been part of the CPS since its inception in the 1940's [10, p. 10]. Basically, the adjustment has only "corrected" the differential undercoverage of the CPS relative to the previous census. This has usually been done by ratioing the sample estimates to independently derived age-race-sex population totals obtained by carrying forward decennial census population estimates to account for subsequent aging of the population, births, deaths and net (legal) migration.

Some consideration has, of course, been given before today's session [e.g., 12-13] to what might happen if the survey were adjusted to "true" and not just census-level population totals. Our approach differs from these previous efforts in scope but not in purpose. We have, as a result of the 1973 Exact Match Study, much more information with which to attempt adjustments. Our principal goal is still, however, to examine the sensitivity of the survey estimates to the problem of coverage, not to make a definitive statement on what the CPS coverage adjustment should be.

Three alternative survey estimates are compared in the remainder of this paper. A brief definition of each of these is given below: 4/

(a) Initial.--This is the survey estimate before any adjustment for coverage. It is also known as the First Stage weight because it consists of all the estimation steps in the CPS up to and including the application of the first stage factors in effect for March 1973.

(b) Standard.--This is the usual March Supplement estimate. It is obtained by inflating the first stage weighted sample results to the census-level age-race-sex population totals discussed earlier. Further adjustments are also made so that husbands and wives living together have the same sampling weight, while at the same time leaving unchanged the estimates for certain labor force categories.

(c) Extended.--This is an estimate obtained by a combination of adjustments designed to yield a complete "correction" for all the March 1973 undercoverage, not just the differential undercoverage relative to the 1970 Census. It was derived as a byproduct of the 1973 Exact Match Study and consists of adjusting the first stage weighted sample to independent population totals based on Jacob Siegel's Preferred Series D population estimates corrected for the 1970 Census undercount, an independent estimate of the total number of U.S. occupied households, and extensive administrative data from social security and tax records for persons eligible for interview in the March CPS.

It should be noted that neither the standard nor extended coverage adjustments make any special allowance for aliens illegally residing in the United States. To the degree that such individuals are not included in Census Bureau estimates of the true population, they have been omitted from consideration. 5/

1.3 Overall CPS population coverage.--Table 1 displays CPS undercoverage rates in March 1973 for all persons and for persons 14 years and older by age, race and sex. The rates express the discrepancy between the initial and standard estimates and the corrected total CPS-eligible population. 6/

(a) Initial.--The March 1973 CPS before adjustment underestimated the population eligible for interview by 9.4 million or 4.4 percent. For adults 14 years or older the undercoverage was proportionately greater, about 5.2 percent. Table 1 shows that the undercoverage rates were much more severe for males of other races (23.2 percent) than they were for white males (4.7 percent). Even so, the absolute number

of white males missed in the survey (3.2 million) far exceeds the number of other males missed (2.1 million). Women, as the table shows, were generally better covered than men; whites better covered than other races. The worst coverage was for males of other races 22 to 39 years of age where over 30 percent were missed.

(b) Standard.--The standard Census Bureau coverage adjustment (the March supplement weighting) reduces the amount of undercoverage quite substantially in nearly every age-race-sex group. Differential undercoverage still exists, however, and therefore could have an impact on estimates of characteristics which vary greatly from one group to another.

Table 1.--Undercoverage Rates by Age, Race, and Sex Before Adjustment and After Standard Adjustment

(Percent of corrected total CPS-eligible population)

Age and Sex	All Races		White		Other Races	
	Before	After Standard	Before	After Standard	Before	After Standard
OVERALL.....	4.4	2.6	2.7	1.9	15.5	7.1
Under 14 years old....	2.2	2.3	0.3	1.4	11.9	6.9
14 years or older.....	5.2	2.7	3.5	2.0	17.2	7.2
MALES						
14 years or older, total.....	6.9	3.7	4.7	2.8	23.2	10.1
14 to 21 years.....	3.6	2.0	1.5	1.6	15.9	4.0
22 to 39 years.....	11.6	5.0	8.9	3.8	30.7	13.2
40 to 64 years.....	5.6	4.2	3.3	3.0	24.4	14.2
65 years or older...	3.2	1.4	2.6	1.7	8.3	-1.5 *
FEMALES						
14 years or older, total.....	3.5	1.7	2.4	1.3	11.8	4.7
14 to 21 years.....	2.1	1.4	0.8	1.1	9.3	3.2
22 to 39 years.....	4.6	2.2	3.4	1.8	12.9	5.1
40 to 64 years.....	2.8	1.2	1.6	0.6	12.0	5.5
65 years or older...	5.0	2.4	4.0	2.2	15.0	4.5

*In this case, the standard estimate exceeded the corrected population total.

It might be good to mention one more thing about the March 1973 CPS population undercoverage before going on to look at the impact of alternative adjustments on income distribution statistics. Insofar as we can tell, the basic demographic dimensions of the missed groups are roughly the same as those which have been observed historically [4, Chart 3]. The only difference of any importance is that the overall coverage is slightly better in 1973 than that in more recent years. This implies, among other things, that any sensitivity we might observe would in all probability be greater if we were doing the same study with, say, the March 1977 CPS.

2. INCOME STATISTICS FOR PERSONS

A considerable body of conjecture exists about the socio-economic characteristics of persons not covered by the decennial censuses or the CPS. For example, the supposition has been advanced earlier at this session [4] that persons missed in enumerated CPS households tend to have smaller incomes on the average (i.e., are poorer) than covered persons of the same sex, race and age. They may also have a weaker attachment to the employed labor force; that is, be more often unemployed.

There is some evidence from the 1970 Census-CPS Match Study supporting the hypothesis that a household's coverage in the Census was directly related to the amount of income received by its members. Median family income for persons missed in the 1970 Census was 73 percent of the median family income of the entire population [5, p. 8]. This pattern of difference applies, however, only to white families; no such relationship emerged for variation of coverage with respect to income among families of other races.

A natural question to ask of the CPS coverage adjustments presented in this paper is whether or not they yield results which conform to working hypotheses such as those just mentioned. In the remainder of this section we will try to provide some answers to this question for statistics on the income of persons.

2.1 Income recipiency.--The number of persons 14 years or older reporting money income in the CPS rose by 7.4 million from the initial to the extended estimates and by 3.5 million from the initial to the standard estimates. The percentage increases of 6.2 percent and 3.0 percent, respectively, are slightly higher than the corresponding percentage increases for all persons. The overall income recipiency rate, consequently, has increased from 78.8 percent at the initial stage to 79.4 percent at the extended stage. Moreover, as table 2 shows, the slight increase in income-recipiency rates is broadly based. All four age groups and all four race-sex groups show increases from the initial to the extended estimates.

Table 2.--Income Recipiency Rates for Persons 14 Years or Older

(In percent)

Item	Initial	Standard	Extended
Overall.....	78.8	79.0	79.4
AGE			
14 to 24 years.....	66.6	67.1	68.0
25 to 44 years.....	80.9	81.2	81.5
45 to 64 years.....	82.1	82.1	82.3
65 years or older.....	91.3	91.3	91.8
RACE AND SEX			
White males.....	92.5	92.6	92.7
White females.....	66.7	66.7	67.5
Males of other races...	84.1	84.4	84.9
Females of other races.	71.5	72.1	72.1

These results are perhaps inconsistent with the working hypothesis just discussed that persons missed in the survey may have a weaker attachment to the employed labor force than do covered persons and hence smaller or no income from earnings.

Mean total money income reported in the survey falls, however, as a result of both the standard and extended coverage adjustments. (See Table 3.) This result is clearly consistent with at least the second part of our hypothesis since it implies a mean income among the persons missed in the survey which is smaller than that for covered persons.

Table 3.--Mean Income Amounts for Persons with Income 14 Years or Older

(In dollars)

Item	Initial	Standard	Extended
Overall.....	6,398	6,376	6,289
AGE			
14 to 24 years.....	2,798	2,847	2,837
25 to 44 years.....	8,223	8,182	8,089
45 to 64 years.....	8,405	8,377	8,246
65 years or older.....	3,941	3,932	3,909
RACE AND SEX			
White males.....	9,001	8,980	8,841
White females.....	3,608	3,609	3,616
Males of other races....	5,744	5,615	5,653
Females of other races..	3,349	3,362	3,336

Why, then, should income-recipiency rates rise, given that mean money income falls (on the average) and given what we believe to be the nature of CPS undercoverage? Two explanations can be advanced. The first is that while a weak attachment to the employed labor force clearly implies low earnings, it does not necessarily imply no earnings and certainly not a total lack of income.

Another way in which we have to modify the hypothesis relates to certain dependent groups often without income of their own: wives living with their husbands (especially if they have young children) and teenage children still living with their parents and younger siblings. These groups tend to be better covered than persons of the same age, race or sex who are living in different circumstances. The effect of trying to account for this difference (as is done in the extended procedure) results in decreasing the original weight of these groups relative to that of groups more likely to have some income. (For white women and persons 14 to 24 this effect is so strong that it not only increases recipiency rates, it also increases mean incomes for persons with income.)

2.2 Income aggregates.--Table 4 shows the Bureau of Economic Analysis (BEA) 1972 benchmarks for each type of money income collected in the March 1973 CPS. Also shown is the percentage of each income type obtained from the CPS at the initial, standard, and extended stages of estimation.

Table 4.--March 1973 CPS Aggregrate Income by Type as Percent of BEA Benchmark Before and After Coverage Adjustment

Type of Income	Revised BEA Benchmark (in billions of dollars)	CPS as a percent of Benchmark		
		Initial	Standard	Extended
Total money income.....	867.0	87.5	89.7	91.3
Wages and salaries.....	619.9	94.5	97.1	99.0
Nonfarm self-employment.	56.5	94.2	95.9	95.4
Farm self-employment....	16.3	64.8	65.3	65.5
Social security 1/.....	39.8	91.1	93.1	94.9
Property income.....	75.0	44.2	45.1	45.8
Public assistance.....	10.9	67.9	70.8	72.9
Other transfers.....	27.2	66.3	68.1	70.6
Other income.....	21.4	64.1	65.6	66.6

1/ Includes Railroad Retirement.

Increasing the number of income recipients will, of course, raise the aggregate amount of money income estimated in the CPS. Note, for example, that the standard adjustment raises the aggregate amount of money income reported from 87.5 percent of the benchmark to 89.7 percent. The extended adjustment lifts the aggregate amount reported to 91.3 percent. The shortfall in the CPS reporting of personal income is reduced by 30 percent after an extended coverage adjustment, 18 percent after the standard adjustment.

The underreporting and nonreporting of income in the survey [16-17] are perhaps the chief causes of the remaining BEA-CPS differences. A full discussion of how those problems occur is beyond the scope of this paper. Other papers [e.g., 18-20] from the Exact Match Study have addressed this question and interested readers may wish to consult them for further information.

2.3 Income distributions for missed persons by race.--Table 5 compares the percentage distribution of CPS total money income for persons with income age 14 or older covered in the survey at the initial stage with that for persons missed in the survey as estimated by the standard and extended coverage adjustments. The results, at least for the extended estimates, are entirely consistent with the findings of Siegel presented earlier [5] for the 1970 Census.

Focusing on the distribution of white persons, it is clear that the extended coverage adjustment picks up persons who tend to have much lower incomes than covered persons. This is not the case for the standard adjustment. The mean income for white persons with income imputed as missed by the standard adjustment was $6,506. This compares to a mean of $6,620 for covered persons with income and a mean of $4,513 for persons imputed as missed by the extended adjustment. (See table 5.)

Unlike those for white persons, the income distributions imputed to covered and missed persons of other races are more nearly identical. This is true of both the standard and extended adjustments and is entirely in keeping with [5].

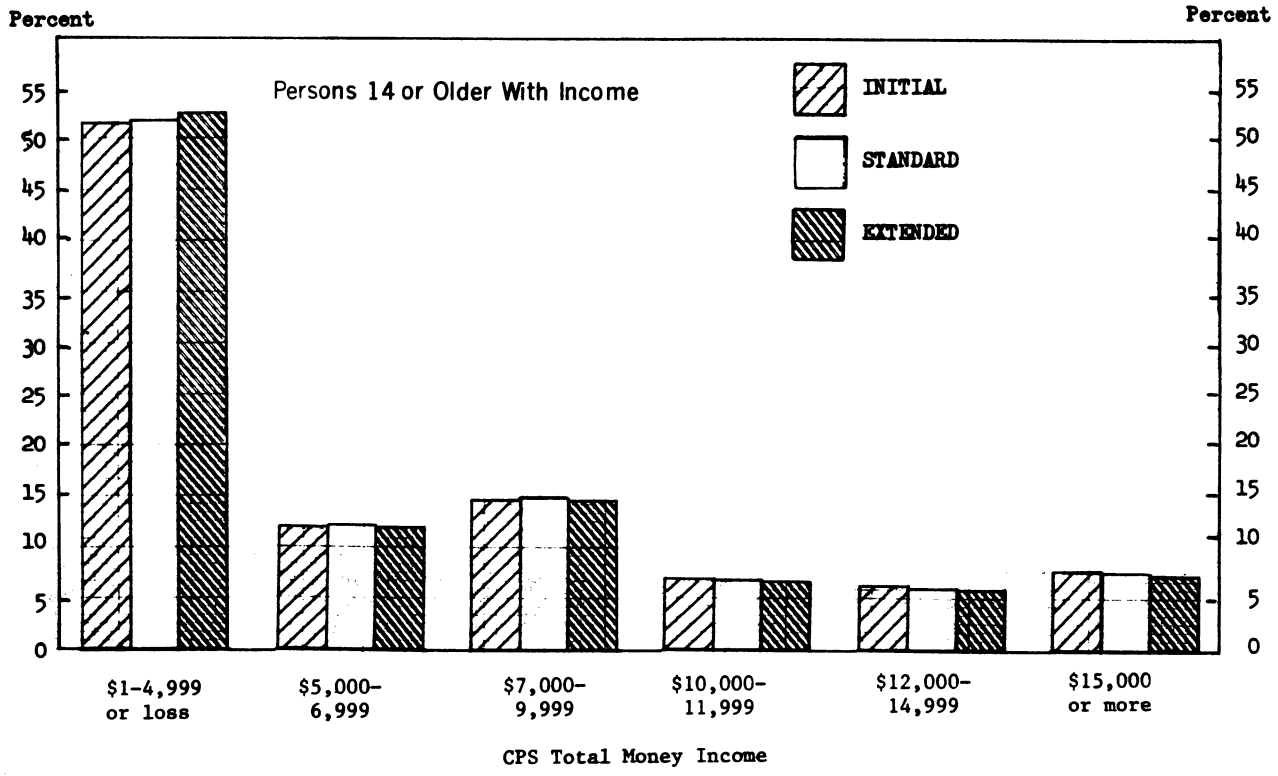
Table 5.--Income Size Distribution of Persons with Income 14 Years or Older, Covered and Missed, by Race

CPS Total Money Income	All races			Whites			Other races		
	Initial Covered Persons	Missed Persons		Initial Covered Persons	Missed Persons		Initial Covered Persons	Missed Persons	
		Standard	Extended		Standard	Extended		Standard	Extended
Total number (millions)..	118.5	3.5	7.4	106.2	1.9	4.5	12.3	1.6	2.9
PERCENT.....	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$1-$4,999 or loss..	51.9	54.4	64.4	50.5	45.6	64.6	64.1	64.6	64.2
$5,000-$6,999.....	12.2	15.5	12.4	12.1	16.5	12.8	12.9	14.3	11.6
$7,000-$9,999... ..	14.3	16.1	14.2	14.4	19.1	13.9	13.3	12.7	14.7
$10,000-$11,999....	7.2	6.1	4.5	7.4	7.4	4.4	4.8	4.6	4.6
$12,000-$14,999....	6.3	4.1	2.6	6.7	5.6	2.4	2.7	2.1	2.9
$15,000 or more....	8.1	3.8	1.9	8.8	5.7	1.9	2.1	1.7	2.0
Mean income (in dollars).....	6,398	5,631	4,526	6,620	6,506	4,513	4,489	4,376	4,530

2.4 Overall distributional impact.--What impact do the alternative coverage adjustments have on the entire distribution of personal money income? Very little, it would seem, given the small size of the adjustment in relative terms. (See figure 1.) Differences between the initial, standard, and extended distributions are extremely small. The differences do, however, follow the hypothesized [5] pattern: Both of the adjusted distributions are shifted slightly to the left of the initial estimates. The extended adjustment, again, shows the larger change, in effect shifting weight directly from the $10,000 and above classes down to the lowest income class. There is an increase in the under $5,000 group of 0.8 percentage points from the initial (51.9 percent) to the extended (52.7 percent); all of this increase is compensated for by declines in the size classes above $10,000, there being no change in the proportion of individuals with incomes of $5,000 to $9,999.

Figure 1

Distribution of CPS Total Money Income for 1972 Before and After Coverage Adjustments



Still another way to look at the distributional differences between the three estimates is to examine the following selected income percentile points.

Percentile Points	Income at Percentile (In dollars)		
	Initial	Standard	Extended
20th.....	1,359	1,370	1,352
50th.....	4,710	4,703	4,609
80th.....	10,363	10,315	10,193
95th.....	17,566	17,476	17,278

Again we see the downward shift in the income distribution after adjusting for coverage. For the standard procedure it is fairly slight ($7, for example, at the median or 50th percentile); for the extended adjustment, the shift is somewhat larger (from $4,710 to $4,609 at the median).

2.5 Measures of Distributional Inequality.--Mixed results have been obtained with respect to some measures of distributional inequality. The overall Gini concentration ratio or coefficient of inequality declines slightly from the initial to the standard and extended stages, from 0.5042 to 0.5029 and 0.5036, respectively. Since the Lorenz curves which underlie those Gini coefficients do not intersect, it can be said unambiguously that measured inequality in the total distribution has been reduced, however slightly, by each coverage adjustment. 7/

A more revealing measure of changes in the relative distribution, (that is, of changes in the shares of various quantiles relative to one another), is the mean income of any quantile divided by the mean income of the distribution as a whole [22, p. 247]. However, even this more sensitive "relative mean income" measure registers very little change as a result of coverage adjustments to the CPS. The mean income of the bottom quintile has increased 2.6 percent relative to the mean income of the distribution as a whole after the extended coverage adjustment; all other quintile share changes are less than 1 percent of the new mean.

3. INCOME STATISTICS FOR FAMILIES AND UNRELATED INDIVIDUALS

In this section we will continue our analysis of the impact of coverage adjustments on income statistics. Attention will be focused now on consumer units (families and unrelated individuals) rather than on persons.

3.1 Number of families and unrelated individuals.--Adjusting for net CPS undercoverage errors has a marked impact on the relative number of families and unrelated individuals. (See table 6.) At the initial stage, unrelated individuals comprise 16.1 million or 23.2 percent of the 69.4 million consumer units. The group rose in importance to 23.6 percent after the standard coverage adjustment; and, after the extended coverage adjustment, unrelated individuals represent 25.2 percent of the total. The difference for unrelated individuals between the standard and extended estimates (about 1.5 million) is even more striking when one notes

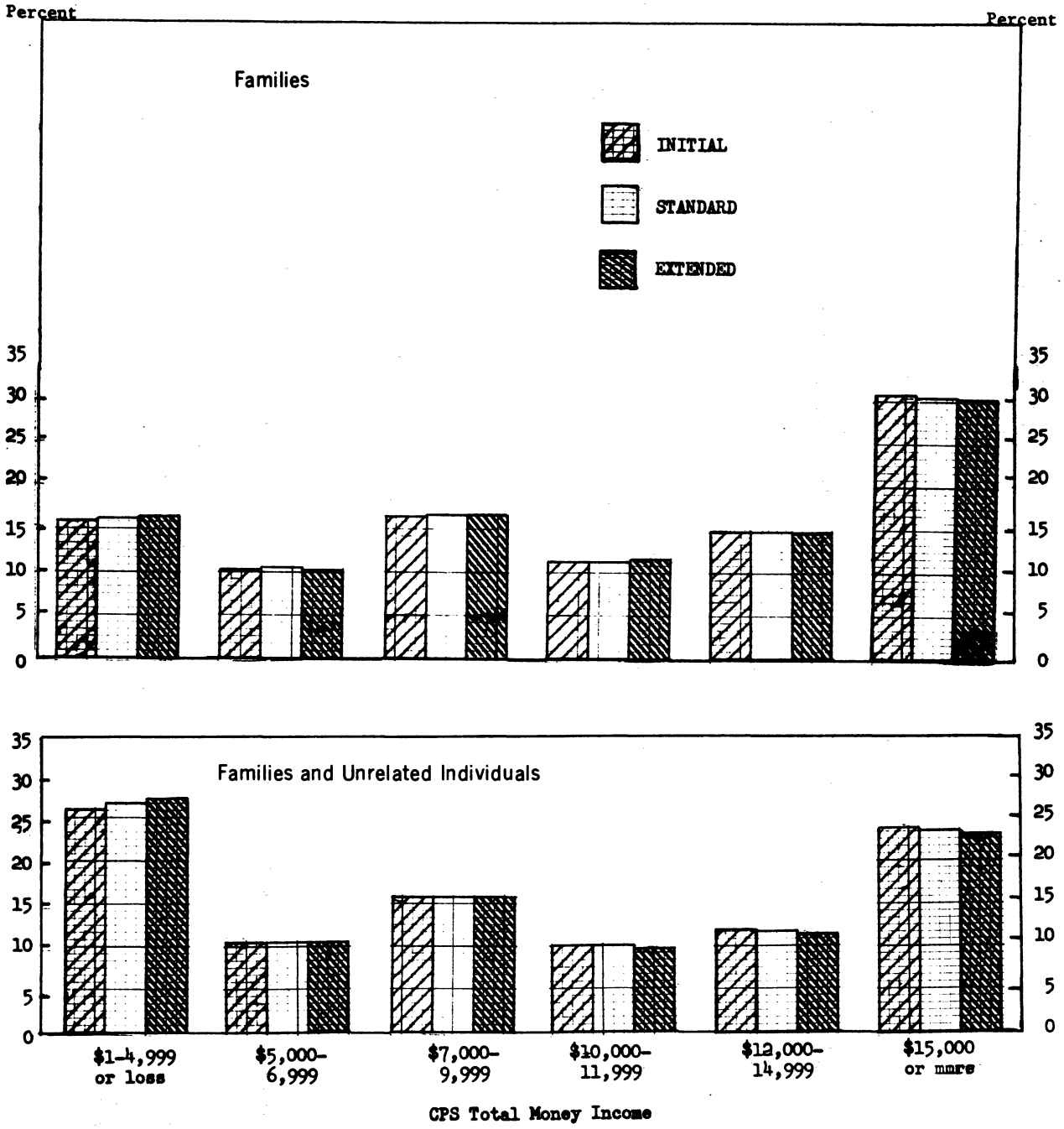
Table 6.--Number of Families and Unrelated Individuals by Type of Estimate, Age, Race and Sex

(In millions)

Item	Families and unrelated individuals			Families			Unrelated individuals		
	Initial	Standard	Extended	Initial	Standard	Extended	Initial	Standard	Extended
Total.....	69.4	71.2	72.7	53.3	54.4	54.4	16.1	16.8	18.3
AGE OF HEAD									
14 to 24 years.....	6.3	6.7	6.9	4.0	4.2	4.1	2.3	2.5	2.8
25 to 44 years.....	25.6	26.2	27.0	22.3	22.7	22.9	3.3	3.6	4.1
45 to 64 years.....	24.0	24.5	24.9	19.6	19.9	19.9	4.4	4.6	5.0
65 years or more.....	13.4	13.8	13.9	7.4	7.6	7.5	6.0	6.2	6.4
RACE AND SEX OF HEAD									
White males.....	48.5	49.3	49.4	43.2	43.8	43.4	5.3	5.5	6.0
White females.....	13.5	13.7	14.2	4.6	4.7	4.7	8.8	9.0	9.5
Males of other races.....	4.7	5.1	5.9	3.7	4.0	4.3	1.0	1.2	1.6
Females of other races...	2.8	3.1	3.2	1.8	1.9	1.9	1.0	1.1	1.3

Figure 2

Distribution of CPS Total Money Income for 1972 Before and After Coverage Adjustments



that the number of families is virtually identical (at 54.4 million) for both procedures.

Although the total number of families did not change between the standard and extended coverage adjustments, the race-sex composition of family heads changed significantly. Particularly noteworthy is that families headed by males of other races jumped 16.2 percent from the initial to the extended stage and 7.5 percent from the standard to the extended stage. Large relative shifts also occurred in the race-sex and age composition of unrelated individuals as a result of the coverage adjustments.

3.2 Distributional impact.--Figure 2 shows the CPS money income distribution before and after adjustment. This is done separately for families, and families and unrelated individuals combined. What should be noted about these size distributions of income is, first, that the initial, standard and extended are

quite close; the second thing to notice is that, as with persons, both coverage adjustments introduce a very small yet persistent downward shift in income.

Perhaps an even better picture of what is happening emerges when we look at median income. For families, median income falls by about 1 percent between the initial and extended estimates, from $11,101 to $10,990 respectively. The standard estimate yields a median family income which, at $11,045, is about midway between the other two.

For unrelated individuals, there is virtually no significant change in the medians. The initial median is $3,526 with the standard ($3,538) and extended ($3,540) being slightly larger.

For families and unrelated individuals combined we see the largest differences between the medians. There is a drop from the initial estimate of $9,304 to $9,225 for the standard and $9,074 for the extended. The reason for larger declines in the combined distribution than in the components is due to the increase in importance of unrelated individuals commented on earlier.

With respect to distributional inequality, there is a tendency for the coverage adjustments to increase inequality among consumer units, in contrast to the decrease among persons. This is especially the case for families and unrelated individuals combined where the Gini ratios rise from .4085 for the initial estimate to .4094 and .4121 for the standard and extended estimates, respectively.

3.3 Some limitations on family data.--A problem exists with the current family estimates in that we are following the standard Census procedure of using the weight of the family head (primary or secondary) in deriving the estimates. Neither the standard nor extended coverage adjustment reflect the increasing probability of a family not being completely enumerated as household size increases.

By shifting weight from smaller-sized to larger-sized families in order to properly reflect the family size distribution, the income distribution will probably be shifted slightly upward. Such an elaboration on the extended coverage adjustment is currently being tested.

The total impact of such a change is expected to move the extended coverage-adjusted family distribution into closer alignment with the already published standard family distribution. Differences caused by the increased importance of unrelated individuals or certain types of families (e.g., those of other races) are expected to remain; however, distributional shifts within each family type will diminish, becoming even less significant than they are now.

4. SELECTED POPULATION CHARACTERISTICS AND LABOR FORCE STATISTICS

CPS characteristics, other than income, are also sensitive to the nature of the coverage adjustment. This section briefly examines the impact of alternative adjustments on labor force data, poverty rates, educational attainment, and residence statistics.

4.1 Labor force and unemployment rates.--The effect of adding about 8.2 million persons age 14 and older to the CPS population (4.0 million more than the standard CPS coverage adjustment) naturally tends to increase aggregate labor force totals. Table 7 shows that, relative to the published estimate [23], the extended coverage adjustment increased the size of the labor force by three million persons or 3.4 percent. While the estimated employed segment of the labor force was increased 3.0 percent, the estimated unemployed segment increased 11.6 percent. This meant that we imputed an unemployment rate of 17.6 percent to the population not covered by the standard CPS coverage adjustment. The overall impact on the measured unemployment rate was to raise it from 5.17 percent to 5.57 percent.

Table 7.--Labor Force Estimates for March 1973 Before and After Coverage Adjustment

(Numbers in millions)

Labor force category (16 years and older)	Type of estimate		
	Initial	Standard 1/	Extended
Total civilian labor force.....	84.7	87.3	90.3
Employed.....	80.3	82.8	85.3
Unemployed.....	4.5	4.5	5.0
Unemployment rate (percent).....	5.3	5.2	5.6

1/ Obtained from [23].

The unemployment rate grows from the standard estimate to the extended one in part simply because we have given certain groups (e.g., males of other races) more importance. Some increase in unemployment rates also occurred, though, for each race-sex group separately.

Johnston and Wetzel in a 1969 paper examined the possible impact of undercoverage on unemployment rates [13]. Their results are sharply different from ours. They prepared two estimates of the unemployed to look at the problem. The first was under a "comparability" assumption. Missed persons were assumed to have the same labor force characteristics as their peers (sex, race, and age cohort). The second was a "poverty neighborhood" assumption. This procedure assumed that the missed population had the same labor force characteristics as its peers in poverty neighborhoods. Under both assumptions the published unemployment rate for 1967 was raised only from 3.8 percent to 3.9 percent.

Differences in the results of the two studies

arise for a number of reasons, including changes in the nature of the CPS, demographic changes in the labor force and in the undercoverage itself. There is also a deficiency in the Johnston and Wetzel methodology which may account for the greater insensitivity they observed. They did not look at the total CPS coverage problem, only that portion of the shortfall which is accounted for by bringing the CPS up to undercount-corrected totals after it had already been adjusted to census-level population estimates.

4.2 Poverty rates.--Brief mention needs to be made of the sensitivity of poverty estimates to alternative coverage adjustments. The following summary comparison in table 8 may aid in this endeavor.

Table 8.--1972 Poverty Estimates Before and After Adjustment

(Numbers in millions)

Persons 14 Years or Older	Initial	Standard	Extended
Total	150.5	154.5	158.6*
Poor	15.8	16.4	17.3
Nonpoor	134.7	138.0	141.3
Poverty rate (in percent):			
Overall	10.5	10.6	10.9
Imputed to missed	**	15.5	21.2

*This is slightly smaller than the corrected CPS-eligible population 14 or older of 158.7 because the extended adjustment does not force exact agreement with all population totals [11].

**Not applicable.

In keeping with the changes in the overall income distribution we found a slight apparent increase in poverty for all persons 14 or older from the initial (10.5 percent) to the standard (10.6 percent) and extended (10.9 percent). Poverty is a family characteristic and hence would be affected by our failure, so far, to correct the standard or extended estimates for within family undercoverage. As a result, it is likely that the above difference between the initial and adjusted estimates overstates somewhat the impact of the problem of undercoverage on poverty rates.

4.3 Other selected characteristics.--To complete this paper's brief discussion of the impact of alternative coverage adjustments, we will now look at changes which take place in statistics by Census Region, metropolitan residence and educational attainment. Some examination of table 9 will show that basically both the standard and extended adjustments are imputing missed persons in a way which would be roughly appropriate if the pattern of undercoverage described in [5] for the 1970 Census were applicable to the March 1973 CPS. In particular, missed persons are more likely to have only an elementary education and to live in the South. The central city population estimates experience the largest absolute and proportionate increases in the residence category as a result of the coverage adjustments. This is mainly due, however, to the heavy concentration of persons of other races in the central cities. The extended estimate, as usual, shows more of a shift than the standard but the direction is generally the same.

Table 9.--Selected Characteristics by Type of Estimate

(Numbers in millions)

Selected Characteristic	Initial	Standard	Extended
Persons 14 years or older by Census region			
Northeast	36.3	37.2	37.8
North Central	41.2	42.2	43.4
South	46.8	48.2	49.8
West	26.2	27.0	27.6
Persons 14 years or older by residence			
In metropolitan areas:			
Central city	45.6	47.4	48.7
Suburban fringe	57.3	58.6	59.7
Outside metropolitan areas	47.6	48.5	50.2
Persons 25 years or older by educational attainment			
Elementary	26.2	26.9	27.9
High school	57.5	58.9	60.1
College	26.4	27.1	27.6

5. CONCLUSION

In this paper we have done some further exploration of the impact of coverage adjustments on income and other socio-economic characteristics. All of the earlier CPS studies of this type [6, 12, 13, 24], and ours as well, show that such adjustments have their major effect on aggregates, with the effects on overall percentages and rates being much less pronounced.

Especially important perhaps for users of CPS income data is the extent to which aggregate income is increased when coverage adjustments are made. We have seen, for example, that some 30 percent of the CPS understatement of income is eliminated by employing the extended coverage adjustment.

The income size distributions of both persons and families were much less affected by our adjustments. Even so, our results demonstrate that CPS coverage problems significantly limit the survey's usefulness in studying economic well-being.

We would like to urge, along with Bateman [3], that additional research be undertaken to improve the coverage adjustment procedures now being employed in the Current Population Survey. Certainly, as users of the CPS, we will be continuing our own studies in this area.

ACKNOWLEDGEMENT AND FOOTNOTES

The authors owe thanks to a great many people for contributing to the work being presented in this paper. Jean Salter at the Bureau of Economic Analysis (BEA) prepared the Income Benchmark revisions used in table 4. H. Lock Oh, Linda DelBene and Faye Aziz at Social Security (SSA) made major contributions to the extensive background material [11, 14] provided as handouts at the session.

Special thanks also must be given to the other participants at this session for their

very helpful comments. Editorial and other assistance was provided by Wendy Alvey, Ben Bridges and Denton Vaughan at SSA and Gordon Green at the Census Bureau. The typing was done at SSA by Rubye Ellis and Helen Kearney.

1/ The 1973 CPS-IRS-SSA Exact Match Study has been the subject of numerous papers at previous American Statistical Association meetings in 1974, 1975 and 1976. For further details on its goals and content, see, for example, [1] or [2].

2/ Derived from the extended coverage adjustment. It should be noted that we imputed a smaller average household size to missed households than to enumerated ones. This does not seem to be consistent with the findings from coverage checks done in connection with the 1970 Census [5]. However, the two results may be reconcilable if one takes into account the impact of the (uncorrected) household size bias that is typical of CPS noninterviews [6].

3/ The CPS estimate of total occupied units before any coverage adjustment was 66.7 million. The total number of occupied units was estimated to be 69.2 million [11, pp. 30-32].

4/ In the companion paper [11] provided as a handout at the session each of these estimates was described in greater detail than space will permit here. It should also be mentioned that the extended estimator is the average of two different approaches to the correction problem on which results are available separately [14].

5/ For some preliminary estimates of the number of illegal aliens not included in Census Bureau population totals, see [15].

6/ The starting point in developing our estimates of the CPS eligible population was Siegel's Preferred 1970 undercount corrected population totals aged to April 1, 1973. To these an adjustment was then made to exclude the institutional population and that portion of the Armed Forces not eligible for interview. (See [11, pp. 16-18] for full details.)

7/ All percentile estimates, Gini ratios and income shares shown in this paper were prepared from the grouped data in [14] using the procedures discussed in [21].

REFERENCES

[1] "The Role of the Social Security Number in Matching Administrative and Survey Records," 1974 American Statistical Association Proceedings, Social Statistics Section, pp. 126-156. See also, "The Reconciliation of Survey and Administrative Income Distribution Statistics through Data Linkage," 1975 American Statistical Association Proceedings, Social Statistics Section, pp. 119-158.

[2] U.S. Social Security Administration, Studies from Interagency Data Linkages (especially Report No. 4).

[3] Bateman, D.V., "Analysis of Census Bureau National Housing Inventory Estimates," 1977 American Statistical Association Proceedings, Social Statistics Section.

[4] Korns, A., "Coverage Issues raised by Comparisons between CPS and Establishment Employment," 1977 American Statistical Association Proceedings, Social Statistics Section.

[5] Siegel, J.S., Coverage of population in 1970 Census and some implications for public programs, Series P-23, No. 56, 1975.

[6] Scheuren, F., Kilss, B. and Oh, H.L., Studies from Interagency Data Linkages, Report No. 2, 1973.

[7] Siegel, J.S., Estimates of coverage of population by sex, race and age: demographic analysis, 1970 Census of Population and Housing: Evaluation and research program, PHE(E)-4, 1974.

[8] Montie, I.C. and Schwanz, D.J., "Coverage Improvement in the Annual Housing Survey," 1977 American Statistical Association Proceedings, Social Statistics Section.

[9] Brooks, C. and Bailar, B., "Nonsampling errors in the Current Population Survey as they affect the Employment Statistics" (Unpublished OMB working paper appearing in a condensed form in 1977 American Statistical Association Proceedings, Social Statistics Section).

[10] U.S. Bureau of the Census, Concepts and methods used in labor force statistics derived from the Current Population Survey, Series P-23, No. 62, 1976.

[11] Scheuren, F.J., "Methods of Estimation for the 1973 Exact Match Study" (unpublished working paper distributed with [14] at the session; to appear in the series Studies from Interagency Data Linkages).

[12] Siegel, J. (1968), "Completeness of coverage of the nonwhite population in the 1960 Census and current estimates, and some implications," Social Statistics and the City. Cambridge: Harvard University, pp. 13-54.

[13] Johnston, D. and Wetzel, J. (1969), "Effect of the census undercount on labor force estimates," Special Labor Force Report No. 105. Bureau of Labor Statistics.

[14] Yuskavage, R. and Oh, H.L., "Four alternative Estimates of CPS income data for 1972" (unpublished working paper distributed with [11] at the session).

[15] Lancaster, C. and Scheuren, F., "Counting the Uncountable Illegals: Some initial Statistical Speculations employing Capture-Recapture Techniques," 1977 American Statistical Association Proceedings, Social Statistics Section.

[16] U.S. Bureau of the Census, Current Population Reports, Series P-60, No. 90, pp. 24-25.

[17] Budd, E., Radner, D. and Hinrichs, J., "Size Distribution of Family Personal Income: Methodology and Estimates for 1964," Bureau of Economic Analysis Staff Paper No. 21, 1973.

[18] Herriot, R. and Spiers, E., "Measuring the Impact on Income Statistics of Reporting Differences between the Current Population Survey and Administrative Sources," 1975 American Statistical Association Proceedings, Social Statistics Section.

[19] Kilss, B. and Alvey, W., "Further exploration of CPS-IRS-SSA Wage Reporting Differences," 1976 American Statistical Association Proceedings, Social Statistics Section.

[20] Vaughan, D. and Yuskavage, R., "Investigating Discrepancies between Social Security Administration and Current Population Survey Benefit Data for 1972," 1976 American Statistical Association Proceedings, Social Statistics Section.

[21] Oh, H.L., "Osculatory Interpolation with a Monotonicity Constraint," 1977 American Statistical Association Proceedings, Statistical Computing Section.

[22] Budd, E., "Postwar Changes in the Size Distribution of Income in the U.S.," American Economic Review Proceedings, May 1970, pp. 247-260.

[23] U.S. Bureau of Labor Statistics, Employment and Earnings, April 1973, p. 23.

[24] Vaughan, D. and Ireland, C.T., "Adjusting for coverage errors in the March 1973 Current Population Survey," 1975 American Statistical Association Proceedings, Social Statistics Section.

DISCUSSION

Jeffrey S. Passel, Population Division, U.S. Bureau of the Census

The authors' technical abilities and knowledge of their data sets as displayed in these papers do not require comment or criticism. Consequently, I will not discuss specific issues addressed in the papers but rather I will comment on the coverage problem in general, with illustrations from these papers.

The problem of measuring coverage of a census or survey, such as the CPS, is very difficult indeed. It is difficult enough to try to measure the coverage of population in age-race-sex groups for a census but when we attempt to measure coverage for geographic areas or according to socioeconomic characteristics, the solution of the problems involved can become practically impossible. As the three papers presented here have shown, the problems are not insoluble, but may require numerous assumptions and may still be very difficult.

Estimating coverage involves comparing the count or estimate obtained from the census or survey with an estimate obtained from independent sources. Because the undercount is a residual, the estimation procedure requires highly accurate and precise data for both the count and the independent estimate. When the survey estimate is not precise, it can be almost impossible to measure coverage. For example, if the standard error of the survey estimate of group size is larger than the probable undercount of this group, it may be impossible to derive a sensible coverage estimate. In such cases, we must resort to making plausible assumptions about the data and the resulting undercount estimates (if any) can be highly variable and unreliable.

The quality of data used for the independent estimate is also important. Let me say first that what we are doing is attempting to measure the survey data against a standard which is almost always unknown and, in most cases, unknowable. Sometimes we may feel we know what the true value of an aggregate total is, but seldom do we know with any confidence what the true distribution of a characteristic is. If we did know the true characteristic distribution then the problem of measuring coverage would be trivial. Furthermore, we really wouldn't have to take the survey in many cases, since the independent estimate would suffice. However, when the true distribution of characteristics is unknown, we must resort to inference or general indications from partial estimates of coverage. A good example of this sort of detective work is the paper by Alex Korns.

In attempting to explain anomalies in the CPS employment series, Korns hypothesizes that coverage problems in the CPS could account for the observed irregularities. To bolster his case, he examines alternative explanations which turn out to be lacking for one reason or another. Then, he turns to some information about missed persons in the Census and CPS. From these bits of information, he builds a strong case for coverage problems in CPS being the cause of the anomalies. His conclusions seem correct but his case is basically inferential because the true distribution remains unknown and the information about missed persons is generally sketchy.

The methods for measuring coverage are limited only by the availability of independent data and the ingenuity of the researcher but they fall generally into four categories:

1. Component or demographic analysis involves building an estimate from components (births, deaths, and migration for population estimates), as well as using information regarding known internal regularities in the data, such as sex ratios. Population estimates made this way by Siegel and others in connection with the 1970 Census are employed by all the authors in one way or another. Bateman also refers to the difficulties involved in generating such estimates for housing.

2. Reinterview studies consist of the reenumeration of a sample of households to check their coverage in the census or survey. Reinterview studies generally do not provide good estimates of overall coverage because of problems in obtaining "true" matches and nonmatches and because of the so-called "correlation bias"; that is, the resurvey misses people, too, and these tend to be the same people who were missed originally. However, reinterviews can provide a great deal of information on the components of error. We can generally distinguish underenumeration from overenumeration, misses within covered units from omitted units, errors of omission from reporting errors, and types of persons missed.

   The primary value of reinterview studies is that they can provide a great deal of information about the characteristics of missed persons as well as components of error. Much of the inference in the Korns and Yuskavage-Hirschberg-Scheuren papers is based on such information from reinterviews. Note that a reinterview study may not provide a quantitative estimate of the error, but can obviously still be quite useful.

3. Record checks involve comparison of census or survey records with a list of persons who should be in the census or survey. This list (or lists) is usually a set or sets of administrative records, such as driver's licenses, social security files, etc., or it could be another survey. By using a set of records which are independent of the census or survey, the correlation bias can be greatly reduced. However, the problems of

obtaining true matches and true nonmatches are increased because of differences in format and scope of data. This method also can provide information on components of error and limited information about characteristics of missed persons. The paper by Yuskavage et al is based in part on coverage information obtained from record checks using Social Security and Internal Revenue records.

4. Comparison with administrative aggregates (used by all of the authors) is another general type of method for estimating coverage. Birth records, social security data, Medicare files are examples of the type of data used. The data may refer to the entire population, or more often, specific age-sex segments. In most cases, the administrative data must be adjusted for known classes or omitted persons or for differences in definition of characteristics.

Results from all four basic techniques for estimating coverage can be manipulated with various statistical methods such as regression or contingency table techniques.

Given that we can estimate coverage of censuses and surveys within some range of error, there are some other issues which must be faced in using such estimates. First, let us be sure to note that even though the undercoverage of a survey may be substantial, much of the information obtained may be virtually unaffected by coverage error. One example is the percentage distribution of population into income classes shown in the Yuskavage-Hirschberg-Scheuren paper. Although the income distribution of missed persons is substantially different from that of covered persons, the corrected distribution is quite similar to the uncorrected and some parameters (e.g., the Gini index) are almost identical. Another such example comes from our work at the Census Bureau on estimating the coverage of the population of States. Although the undercount in some States is moderately large, the percentage distribution of the State populations change very little when corrected for undercount. In cases such as these, it is only the differential undercount of income classes or States that change the percentage distribution. Furthermore, it takes a substantial difference to alter the basic distribution more than a very small amount.

Another issue that must be faced in using any coverage estimates is what level of error can be tolerated; in other words, when is it preferable to use the corrected numbers over the uncorrected ones. The most stringent error limitations should be placed on corrections for numbers which are used to disburse funds competitively. If the coverage estimates for such numbers can be in error, then the allocations for some areas based on the "corrected" numbers may be further from the "true" allocation than those based on uncorrected numbers. In such cases, the cause of equity will not be served by using "corrected" numbers. This type of competitive allocation requires coverage estimates of a high degree of accuracy and precision, as well as uniform quality for all areas considered.

For noncompetitive allocations, such as capitation grants, the requirement of uniformity in quality may be relaxed, but the accuracy requirements remain. A lower level of accuracy and precision can be tolerated for coverage estimates used in research. Such estimates can be used to indicate whether or not research results are caused by or altered by coverage errors. A still lower level of accuracy and precision can be tolerated in coverage estimates which are used illustratively. Such estimates can still provide qualitative indications of errors and can be useful as rough guides in the broad interpretation of census or survey data even though they may be somewhat inaccurate or imprecise.

The estimates presented in this session generally fall in the middle category. The results obviously have found research applications, but must be refined for the more demanding uses. In conclusion, I would like to commend the authors for their work. Furthermore, I would like to recommend strongly to users of CPS and census data that they take heed of the findings presented here in the course of their own research.

THE FOUNDATIONS OF EPIDEMIOLOGY: THE FRENCH CONNECTION*

David E. Lilienfeld and Abraham M. Lilienfeld, The Johns Hopkins University

We thought that we would begin our presentation of the historical foundations of both health statistics and epidemiology with a brief summary of the present state of both. Today, statisticians have several tools with which to handle data, some of which are shown in Table 1. These include the normal distribution, relative risk estimators, age-adjusted rates, confidence intervals, the census, experimental designs, and life tables. Yet all of these tools were developed and/or used by 19th century epidemiologists and statisticians during the historical period that I (D.E.L.) refer to as the Greening of Epidemiology.

The seeds of the epidemiological tree were sown by the French, shortly after the French Revolution of 1789 (1). Statisticians are familiar with the works of Laplace and Poisson in the early 1800's. Indeed, in his Philosophical Essay on Probability, Laplace describes our "modern" life table, makes the interesting statement that "A table of mortality is then a table of the probability of human life", and describes an approach to the analysis of competing risks (2).

The French Revolution represented a break with past traditions and customs (1). In medicine, a similar break with the past occurred. Prior to the Revolution, a theory of medicine provided the basis for its practice; afterwards, however, physicians decided to begin with the actual practice of medicine and generalized this into theory. The instrument of this generalization process was statistics, and the key figure in the medical adoption of statistics was the epidemiological pioneer, P.C.-A Louis. Louis, whom we view as one part of the "French Connection", championed what he called the "numerical method". He consistently employed this statistical approach to medicine, using it in 1830 to show the ineffectiveness of bloodletting as a therapeutic agent. Although Cochran has noted that modern experimental designs were used in agricultural research as early as the 1970's, it is interesting to note that Louis described a balanced block design in his book on the evaluation of bloodletting. He stated:

> "To this I reply that the calculus as I employ it, does not efface differences: it supposes them; it limits itself to combining similar unities in order to compare them with parallel unities, these being subjected to somewhat different influences; that is, after all, as has been before remarked, it should sometimes be necessary that facts should be combined which are not strictly similar. The error will be distributed through the different groups or classes of facts, and will be equalized; so that a comparison can be instituted between several groups without altering the result."(3).

Louis was also familiar with the general concept of a prospective study, as shown in his approach to the determination of whether or not phthisis was inherited, stating:

> "...to determine the question satisfactorily, tables of mortality (life tables) would be necessary, comparing an equal number of persons born of phthistical parents with those in an oposite condition.(4).

One of the ideas that dominated 19th-century epidemiology and statistics was the concept of a "law of mortality" or "vitality". These "laws" were not necessarily mathematical, such as Gauss' Law, although they could be; one such law was the doctrine of contagium vivum, the "germ theory" Louis was an advocate of the idea of a law of mortality, as shown in a letter to James Jackson, the father of one of his students, in which he wrote:

> "Think for a moment, sir, of the situation in which we physicians are placed. We have no legislative chambers to enact laws for us. We are our own lawgivers' or rather we must discover the laws on which our profession rests. We must discover the laws and not invent them; for the laws of nature are not to be invented" (5).

Thus, it can be seen that statistics was the quantitative manifestation of the inductive reasoning used by the Parisian school of medicine, following the pattern established by the physicists, who were using the calculus as a quantitative manifestation of the deductive reasoning process for deriving physical laws. Time does not permit a presentation of all of Louis' contributions to statistics and epidemiology.

In 1840, Jules Gavarret, of the Polytechnic School of Paris, a student of Quetelet, published a book entitled "General Principles of Medical Statistics", in which 99% confidence intervals were applied to Louis' data on the effect of bloodletting (6). Gavarret criticized Louis because the latter would not use such confidence intervals. Perhaps Louis understood the distinction between biological and statistical significance, when used in epidemiology better than we do today!

Louis' influence was extended by the work done by his students, among whom were the leaders of mid- and late-19th century epidemiology and statistics. These students form the second part of the French Connection. The American students of Louis are shown in Figure 1. Louis' European students are shown in Figure 2. One of these students, familiar to everyone here, was William Farr.

Farr, one of the leading statisticians of his time, was a titan of mid-19th century epidemiology. As Louis's student in the early 1830's he was instilled with Louis' beliefs in a "law of mortality", having stated:

> "Thus, we learn in the same circumstances the same number of people die at the same ages of the same diseases, year after year; organized bodies governed by laws as fixed as those which govern the stars in their courses" (7).

And, further:

> "The deaths and causes of death are scientific facts which admit of numerical analysis; and science has nothing to offer more inviting than the laws of vitality..." (7).

Throughout his life, Farr searched for these laws by constructing life-tables, etc. For it was Farr who referred to the life table as a "biometer" because of its ability to measure life (1). It appears that either Farr or one of his fellow

actuarians, in the late 1830's or early 1840's, coined the term "force of mortality" as used in life tables. One should be aware that this occurred in the mid-1800's when the physical laws governing electrical forces were being discovered, when physicists were using such terms as the "electromotive force". It is possible that the "force of mortality", was therefore used in a similar manner in various laws of vitality.

(1,7) Farr had an amazing grasp of epidemiologic concepts, as shown in Table 2, and, of course, Farr's guiding philosophy is shown by his statement "The death rate is a fact; anything beyond this is an inference"(7). He worked with one of Louis' other students, Marc d' Espine, to develop the predecessor of today's International Classification of Diseases. Lastly, Farr may be viewed as one of the founders of the English school of statistics. He was active in the Statistical Society of London, predecessor of the present Royal Statistical Society, eventually becoming its president in 1872. He established the British Vital Registration System, including the first turly "modern" national census. He had a strong influence on Francis Galton, apparently interesting Galton in statistics. And, he was an advocate of the rigorous statistical analysis of epidemic data; indeed, in 1854, he noted the relationship of water purity, by water company, with cholera mortality in a statistical manner, laying the foundation for John Snow's classic analysis of the epidemic (Table 3).

William A. Guy, a physician, was another student of Louis' at a time after Farr had already returned to London. He, along with Farr, was one of the founders of the London School of Statistics. He became a Fellow of the Statistical Society of London, and was very active in its activities, serving as editor of its journal and in 1874, as its president. In 1846, Guy became the Dean of the King's College Medical School. He was among the first to note the basis of what today we term a "Berksonian Bias". He stated:

> "There are two questions to which I am not aware that any answer has yet been given; nor has any collection of facts been made with a view to furnish a reply. The first question refers to the class of persons who resort to hospitals; the second to the proportion which that class forms of the population to which they belong" (8).

Guy also noted that the variation of the estimator of the mean in a sample decreased with an increasing sample size (9).

The appropriateness of any application of Gavarret's and other French statistician's theories of probability and statistics to clinical medicine and epidemiology was also investigated by Guy. Unlike today's statisticians and epidemiologists, I guess that Guy did not believe that a statistical <u>theory necessarily</u> had any relevance to the "real world". Indeed, the modern multiple logistic equation, the present fetish of both of both statisticians and epidemiologists, has advanced epidemiological research to such an extent that figure 3 shows our present predicament! In 1855, Guy stated:

> "Gavarret...criticises with some severity the conclusions of Louis respecting pulmonary consumption and fever, on the score of the insufficient number of his facts (collected over 7 years), and insists on applying to those conclusions corrections avowedly drawn from treatises on the doctrine of probabilities. Now, unless I am greatly mistaken, no attempt of any kind has yet been made to show that rules and calculations derived from abstract reasoning upon probabilities, backed by a few experiments on occurrences brought about by what is commonly designated 'chance', are applicable to events of a totally different order, brought about by the operation of the human will or by the multitudeinous external influences which, acting on the human frame, preserve it in health or give rise to the diseases which impair its vigour and utlimately destroy it" (10).

Accordingly, Guy proceeded to perform a series of experiments which would be collectively known today as a "Monte-Carlo" simulation.

He began with a defined population of black and white balls; he proceeded to randomly sample from that population, both with and without replacement. Guy then analyzed the samples and compared the results with the population. He concluded that there was some analogy with Gavarret's theorems and statistical significance. Strangely, even after this simulation, Guy used such methods only once.

Lastly, one of Guy's great contributions to epidemiology was his epidemiological studies of the effect of occupation on health. They read as if they were reports of studies published in one of today's scientific journals. Indeed, in one of Guy's studies, a numerical expression equivalent to the odds ratio was used, seemingly, as an estimate of relative risk, which was well-known to have been used by mid-19th century epidemiologists and statisticians. One other important figure was an actuarian, F.G.P. Neison, a colleague of Guy's occupational studies. Neison was the <u>first</u> person to use a method of standardization for death rates to account for differences in the age distribution of populations (Table 4) (11).

Of course, after Galton, the history of statistics is well-known. But, there are some recently uncovered details such as the relationship of Guy to Newsholme, who wrote a text entitled, "Elements of Vital Statistics" in 1889 (12). The tree of epidemiology has grown quite a bit since it first "greened" in the mid-1800's, when epidemiology was barely distinguishable from statistics. Many new branches have grown on that tree as epidemiology has developed. Yet, unfortunately, as a result of many circumstances, we believe that we, today's statisticians and epidemiologists, are out on a limb of this epidemiological tree, and that that limb is being cut off from the trunk! One reason for this, we believe, is indicated in Figure 4.

Thus, Kendall's question, posed in 1975, of why statistics developed in the way that it did has been partially answered (13). Statistics was the means of generalization in Post-Revolutionary French medicine. There were two approaches to such generalizations, known as laws of mortality.

Louis advocated a discrete approach - an absolute law - and, hence, had little real use for theoretical statistics; Gavarret advocated a stochastic approach - a probabilistic law - and, hence, freely used theoretical statistics. Today, these two approaches are evident in physics in the form of the classic Einstein-Heisenberg controversy, in epidemiology with the "web of causation" and definite specific causes debate and, generally, in deterministic and stochastic equations.

The greatest value of history is the perspective it allows one to view the present, before it, too, becomes history. One lesson that we have learned from our on-going historical excursions is that the basic structure of epidemiology is composed of methods - methods devised by the epidemiologist and the statistician alike. These methods should continue to be developed by both the epidemiologist and the statistician almost hand-in-hand. The inferences derived from any given study can change, but the method used to conduct that study does not. Indeed, one reason why the histories of both epidemiology and health statistics have not yet been written is the over-emphasis on inferences and the lack of attention to methods. For, as Daniel C. Gilman said of the Johns Hopkins University in 1890:

> "Whatever gains we may make in our material condition, whatever limitations are still obvious, let us not forget , my friends, that men and methods make universities, not halls, nor books, nor instruments, important as these are."

So, the same can be said for epidemiology:

Whatever limitations are still obvious, let us not forget that men and methods make epidemiology, not statistical significance levels, nor computers, nor inferences, important as these are.

---

*Partially supported by a grant from the Milbank Memorial Fund.

References

1) Lilienfeld DE: The greening of epidemiology, sanitary physicians and the London Epidemiological Society (1830-1870). Bull Hist Med (In press)
2) Laplace PS: A Philosophical Essay on Probabilities. New York, Dover Publications, 1951
3) Louis PCA: Researches on the Effects of Bloodletting in some Inflammatory Diseases, and on the Influence of Tortarized Antimony and Vesication in Pneumonitis. Translated by C.G. Putham. Boston, Hilliard, Gray and Co. 1836
4) Louis PCA: Review of pathological researches on phthisis. Amer J Med Sci 19:445-449, 1836
5) Bollet AJ: Pierre Louis: The numerical method and the foundation of quantitative medicine. Amer J Med Sci 266:92-101, 1973
6) Gavarret J: Principles Généraux De Statstique Médicale. Paris, Bechet Jeune et Labe, 1840
7) Farr W: Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr. Metushen, NJ, New York Academy of Medicine, 1975
8) Guy WA: On the nature and extent of the benefits conferred by hospitals on the working classes and the poor. J Roy Stat Soc 19:12-27, 1856
9) Guy WA: On the relative value of averages derived from different number of observations. J Roy Stat Soc 13:30-45, 1850
10) Guy WA: On the analogy between the aggregate effects of operation of the human will and the results commonly attributed to chance. J Inst Act 5:315-323, 1855
11) Neison FGP: On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts. J Stat Soc 7:40-68, 1844
12) Newsholme A: Elements of Vital Statistics. New York, The Macmillan Co., 1899
13) Kendall MG: Statisticians - Production and consumption. Amer Statistician 30:49-53,1975

TABLE 1

TOOLS OF THE MODERN HEALTH STATISTICAN AND EPIDEMIOLOGIST

1)	THE NORMAL DISTRIBUTION
2)	RELATIVE RISK ESTIMATORS
3)	AGE-ADJUSTED RATES (MORTALITY, MORBIDITY, ETC.)
4)	CONFIDENCE INTERVALS
5)	THE CENSUS
6)	EXPERIMENTAL DESIGNS (CROSS-OVER, LATIN SQUARE, ETC.)
7)	LIFE TABLES

TABLE 2

Examples of William Farr's Understanding of Epidemiologic Concepts

Epidemiologic concept	Farr's statement
Scope of epidemiology	"The causes that make the rates of mortality vary may be considered under two heads - (1) Causes inherent in the population itself, such, for example, as sex and age. (2) Causes outside the population, such as air, water, food, clothing, dwellings, or such groups of causes as are involved in residence, and relation of the several parts to each other in time and space."
Person-years	"A year of life is the lifetime unit. It is represented by one person living through a year; or by two persons living through half a year."
Relationship of death rate and probability of dying (or living)	"...the rate of mortality serves to give the probability of living a year..."
Standardized mortality rate	"[If] the number of boys under 5 years of age was 147,390; the annual rate of mortality in the healthy districts [the standard population] was .04348;...6367 deaths which would have happened in London... continuing the process... the mortality in London should [be] 15 in 1,000..."
Dose-response effect	"...the effects are in some regulated proportion to the intensity of the causes...
Need for large numbers of population and biological inferences	"...When the number of cases is considerable the relative mortality is most correctly expressed and...slight differences deserve little attention."
Herd immunity	"The small-pox would be...sometimes arrested, by vaccination which protected a part of the population..."
Prevalence = incidence X duration	"...in estimating the prevalence of disease, two things must be distinctly considered; the relative frequency of their attacks, and the relative proportion of sick-time they produce. The first may be determined at once, by a comparison of the number of attacks with the numbers living; the second by enumerating several times the living and the actually sick of each disease, and thence deducing the mean proportion suffering constantly. Time is here taken into account: and the sick-time, if the attacks of two diseases be equal, will vary as their duration varies, and whatever the number of attacks may be, multiplying them by the mean duration of each disease will give the sick-time."
Retrospective and prospective studies	"Is your inquiry to be retrospective or prospective? If the former the replies will be general, vague, and I fear of little value..."

TABLE 3

Mortality from Cholera in Districts Supplied by Water Companies, 1853

Water Companies	Source of Supply	Aggregate of Districts Supplied Chiefly by the Respective Water Companies: Elevation (in feet) above trinity high water mark	Population	Deaths from cholera in 13 weeks ending Nov. 19	Deaths per 100,000 inhabitants
London	-	39	2,362,236	744	30
Hampstead & New River	Springs at Hampstead and Kenwood, two artesian wells and New River	80	166,956	8	5
New River	Chadwell Springs in Hertsforshire, from River Lee, and four wells in Middlesex and Herts	76	634,468	56	9
Grand Junction	Thames, 360 yards above Kew-Bridge	38	109,636	15	15
Chelsea	Thames at Battersea	7	122,147	22	18
Kent	Ravensbourne in Kent	18	134,200	31	23
West Middlesex	Thames at Barnes	72	277,700	89	32
East London	Lee at Lee Bridge	26	434,694	162	37
Lambeth & Southwark	Thames at Thames Ditton and at Battersea	1	346,363	220	64
Southwark	Thames at Battersea	8	118,267	121	102
Southwark & Kent	Thames at Battersea, Ravensbourne in Kent, Ditches and wells	0	17,805	19	107

TABLE 4

CHADWICK'S CALCULATIONS

	BETHNAL GREEN	MARYLEBONE
MEAN AGE AT DEATH (UNADJUSTED FOR AGE)	25.8	29.12

NEISON'S CALCULATIONS

	BETHNAL GREEN	MARYLEBONE
MEAN AGE AT DEATH (ADJUSTED FOR AGE)	25.8	24.52

Figure 1: The influence of P-C. A. Louis on the development of statistics and epidemiology in the nineteenth century in the United States.

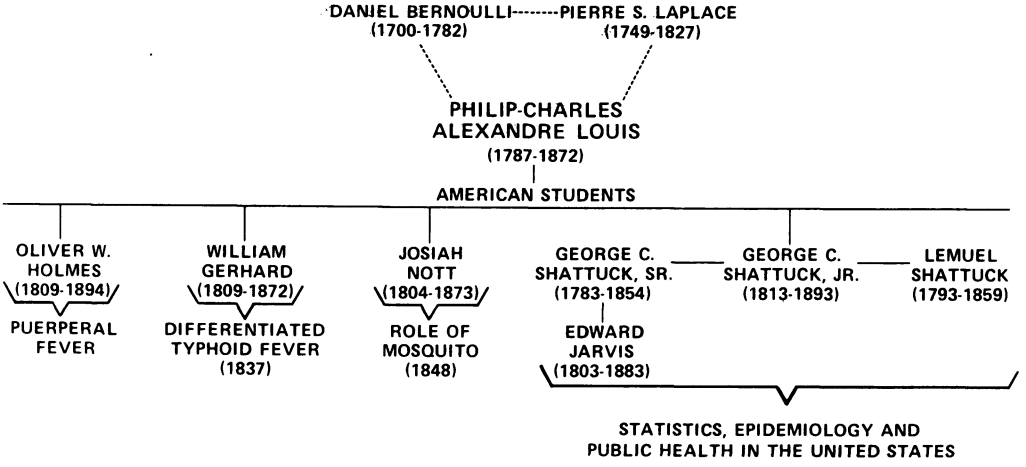
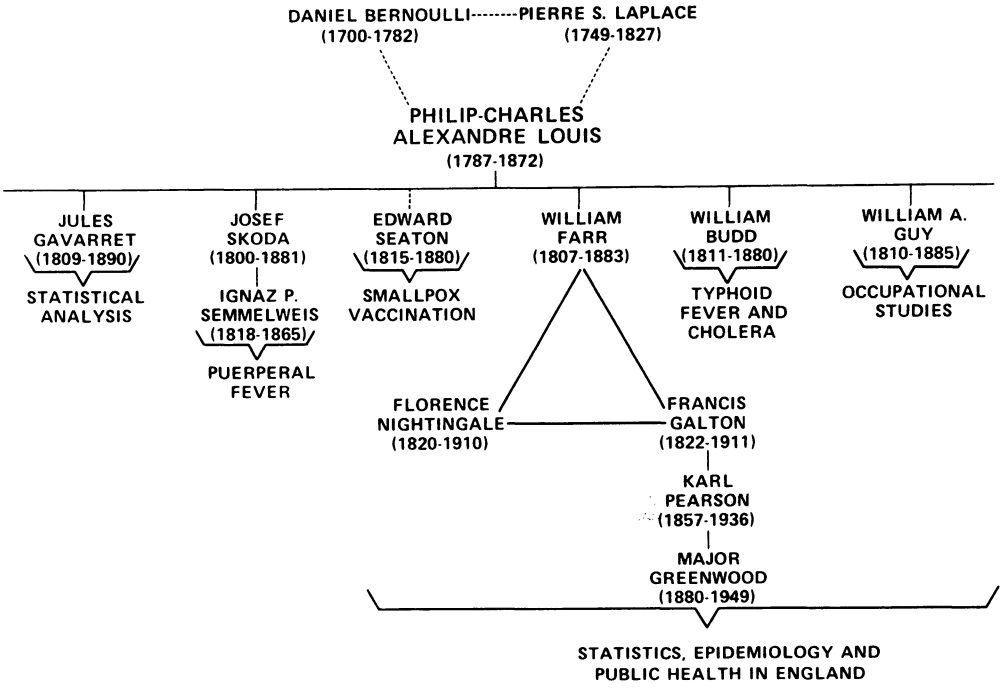
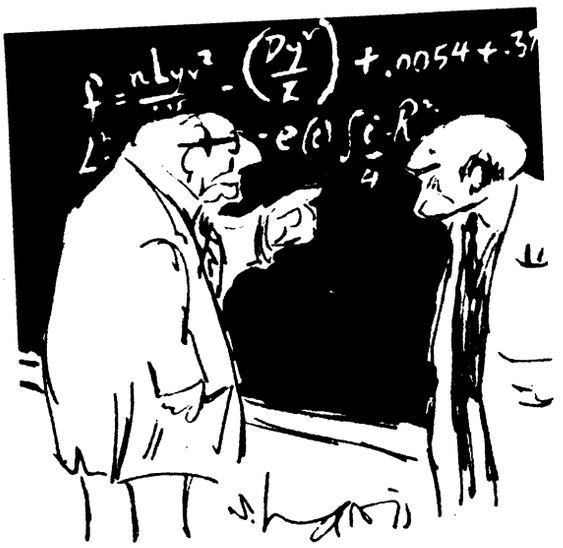


Figure 2: The influence of P-C. A. Louis on the development of statistics and epidemiology in the nineteenth century in England and Europe.





"Does this apply always, sometimes, or never?"

"... and in 1/10,000th of a second, it can compound the programmer's error 87,500 times!"

SPECIAL INVESTIGATIONS ON MORTALITY

Ruy Laurenti, Sao Paulo University

Mortality statistics constitute a fundamental tool in public health and in epidemiology they represent one of the most important sources of information for many kinds of research. According to Mac Mahon,[1] "the introduction of death registration was the foundation of modern epidemiology; it changed the subject from a narrative discipline into a quantitative science".

Historically speaking the vital registrations turned up way before the statistical services and always had, basically, a social meaning, fulfilling therefore a very important function in the community. In law its aim is to reveal certain juridical situations. Thus, its main function is of a juridical nature and consists in registering facts and acts of the civil state allowing the organization and working of the law system ruling the relationship of individuals amongst themselves - the family organization - and its liaisons with the State.[2]

Besides this objective, the statistical function of the vital registrations is to provide data to several sectors such as education, health, economy, industry, commerce, etc.

In the health sector the vital statistics, especially the mortality statistics, are widely used and are of great importance to epidemiological studies.

The primary source of mortality data is the death certificate that is filled out by the physician and thereafter registered, thus starting up the "mortality statistics system". Nevertheless, a great number of physicians ignore the statistical reasons and the uses that the health sector will make of the information they put down in the certificates.

At least in our country, and perhaps in many others also undergoing development, to most of the physicians the death certificate is nothing but a necessary legal document which enables burial and almost always is also necessary for family affairs regarding inheritance matters, social security, etc. The doctors actually forget, or do not even know, the very important statistical function of the death certificate. Here among us, this aspect has not been usually taught in medical schools. This leads to a lack of accuracy, mainly as regards the most important data for epidemiology studies the underlying cause of death.

In the Department of Epidemiology of the Sao Paulo University School of Public Health, some research is under way regarding patterns of mortality for the city of Sao Paulo which we have called "Special Investigations on Mortality". Existing errors due to incorrect filling out of death certificates are dealt with by means of additional information which is obtained by a method which will be commented on later.

In the city of Sao Paulo the death record is virtually complete and 100% of the certificates are filled out by physicians. In cases of violent deaths (or not natural causes) the death certificate is filled out, after autopsy, by the "medico-legista" (coroner); in cases of death due to natural causes and where there was no medical attendance, the death certificate if filled out by a pathologist doctor, after autopsy, on duty at what we call the "Service for the Verification of Death". Almost 20% of all deaths undergo autopsy, including here those where autopsy is asked for in order to clear up diagnosis (mainly in medical school hospitals), those performed by "medico-legista" and those performed at the above mentioned "Service for the Verification of Death".

In this paper some results regarding results obtained through special investigations on mortality for the city of Sao Paulo shall be put forward.

The first research in which we took part, that started up a whole line of investigation at the Department of Epidemiology of the Sao Paulo University School of Public Health, was the "Inter-American Investigation of Mortality",[3] sponsored by Pan American Health Organization (PAHO). It was an international investigation which lasted two years (1962/1963) and that deeply analysed, according to a standardized methodology, the deaths of adults, 15 to 74 year-old age, in 12 cities, 10 being from South America, one from the United States and one from England. In Sao Paulo a sample of 4,361 deaths (one out of six) was investigated and for each case an interview at the home of the deceased was undertaken where besides other information, the place or places where the deceased had received medical attention was obtained. Afterwards, information was collected from doctors, hospital records, necropsy records and from whatever documents related to the case. Thus, it was possible to know the true underlying cause of the death and characterize the patterns of mortality in accordance to several aspects. The results were published by Puffer & Griffith[3]. In 1532 death certificates (35.4% of the total 4361 cases) it was found that the underlying cause declared by the physician was not the correct one. It is interesting, nevertheless, to notice that there are not huge differences as regards greater groups of causes when considering the original death certificates and the results of the investigation. Thus, for example, there were 818 original cases of malignant neoplasms and at the end there were 822. It isn't that the investigation showed up four cases more than the original ones, but that in 196 out of the 818 cases the underlying cause was not malignant neoplasm, thus remaining 622. In addition to these, 200 others, where the investigation discovered malignant neoplasm to be the underlying cause of death and the death certificate did not mention the fact, were summed up. One can say that there was a numerical compensation in this case, but there certainly was no compensation as far as sex and age groups were concerned.

In Appendix I the original classifications and final assignments to some causes of death at ages

15-74 years in Sao Paulo are exposed.

We also took part in the "Inter-American Investigation of Mortality in Childhood",[4] also sponsored by PAHO, where a similar methodology was employed. Here deaths of under-fives were investigated in 14 selected areas totalling 25 projects. Areas from Latin America, one from the United States and another from Canada were included. In Sao Paulo a sample of 4,312 deaths (1 out of 4.25) which occurred between July 1, 1968 and May 31, 1970 were studied. Interviews at the homes of the deceased were also carried out as well as with physicians and at the hospitals which cared for each child, not only during the disease causing death but also before, throughout other consultations and medical episodes. The domiciliary interview collected a greater quantity of information than the anterior investigation performed for adults. We were able to obtain, therefore, information on family composition, type of housing (number of rooms, water and drainage facilities, toilet), occupation of parents, pregnancy history of the mother, data on parents (age, marital status, education), prenatal care, pregnancy complications, type of pregnancy, birth weight, breast-feeding, weaning food used, medical attention received by the child, clinical history, laboratory tests results, autopsy reports, and other information.

In this investigation not only was the underlying cause of death codified, but the associated causes as well.

As regards the underlying cause of death, some important aspects can be pointed out in the results obtained. For example, in the sample studied in Sao Paulo, the original death certificates informed in 91 cases that the underlying cause of death was measles, whilst the investigation showed that there actually were 156 cases, therefore 1.71 times as much. As regards whooping cough, this relation was 1.84. In reference to perinatal causes the number went up from 1,072 cases to 1,119 the relation being, therefore, 1.04. As to this aspect the difference was very small between the original certificates and the final results of the investigation. But, when the specific causes of death among the perinatal ones were analysed, the differences became more intense, as for example in those cases put down as "Difficult Labor" which rose from 14 cases to 168, twelvefold more. In Appendix 2 the infant deaths from certain perinatal causes as underlying causes based on death certificates and on final assignments, with corresponding ratios, are presented.

Another aspect that was able to have been analysed refers to the multiple causes of death. The classification of causes was based on information obtained in the interviews. The underlying cause was classified according to the definition * and the rules for selection and modification set forth in the International Classification of Diseases (8th Revision). Once the underlying cause was selected and the intermediary and terminal diseases or morbid conditions were established, the contributory causes** were determined.

In this "investigation" the multiple causes of death were classified into 2 main groups: underlying and associated. The latter took in both contributory and consequential causes, that is, those morbid conditions that commonly are included in the train of events enhanced by the underlying cause involving both intermediate and terminal ones. As Puffer & Serrano[4] put it,

> ...from the study of interrelations of causes in the Investigation it appears in reality that the intermediate and terminal causes (complications or consequences) are not the result of the underlying cause alone. Instead, it is the complex of underlying-contributory cause which together give origin to the fatal complications. The implications of this concept are exceedingly important from the preventive view point because measures aimed at preventing the underlying cause are not sufficient if the contributory conditions remain. One example is the common association of nutritional deficiency and infection, which necessitates prevention of both components of the complex, without which the phenomenon of "substitution" may come into play - that is, a death prevented by suppressing one underlying cause (usually an infection) may later occur as the result of another underlying cause (another infection) if the contributory condition (nutritional deficiency) remains".

The "Investigation" showed that nutritional deficiency, as associated cause, exerts an important role in the death of the children from Sao Paulo, mainly when the underlying cause of death is an infectious disease. Thus, the prevention of deaths among children should not be directed solely towards the prevention of infectious diseases, but should take in the infection-malnutrition relationship. In 2,354 deaths due to all causes of under-fives, with the exception of the neonatal deaths, nutritional deficiency showed up as underlying or or associated cause of death in 47.1% of the cases. When the underlying cause was due to infectious disease this value was 59.6%. One must point out, nevertheless, that malnutrition did not come to this amount when the original death certificates were analysed, but to much lower values. In Appendix 3 the results of the "Investigation" are shown as regards nutritional deficiency as an associated cause of death, by some groups of underlying causes, in Sao Paulo and in the Latin-American projects on the whole.

The analysis of mortality due to multiple cause was performed in Sao Paulo the first time in the

---

* Underlying cause is "the disease or injury which initiated the train of morbid events leading directly to death".

** Contributory cause is "any other significant condition which unfavorably influenced the course of the morbid process and this contributed to the fatal outcome, but which was not related to the disease or condition directly causing death".

"Inter-American Investigation of Mortality in Childhood"[4]. The good deal of information this investigation lead up induced us to go into another investigation, studying a sample of deaths of all ages, only those that occurred in hospitals of Sao Paulo (70% of all deaths) from March 1, 1972 to February 28, 1973.[5] Differently from the preceding two investigations, in this one additional data were gathered only from the hospital where the deceased received final medical attention. No domiciliary interviews were performed, neither was any sort of information looked up at other hospitals or medical services at which the deceased might have been seen. One of the main objectives was to see whether only data collected at the hospital during the final episode were good enough to furnish information as good as that gathered through relatives and parents and other medical services where the deceased had been before this final hospitalization. In this type of investigation one verified that as many additional diagnoses are obtained as in the other one. In fact, in the sample of 1,832 deaths, 852 were under-fives and for these there were 2,575 diagnoses, that is, an average of 3.0 diagnoses per case. In the investigation performed before this one and which included interviewing at the home of the deceased and at the other hospitals and mdeical services, for a sample of 4,312 cases, we obtained 12,988 diagnoses, an average therefore of 3.0, as well, per case.

For adults this comparison was not possible, as in the anterior investigation there was no research regarding multiple causes of death in adults. For adults in Appendix 4 the data regarding Ischaemic heart disease and its associations with other causes in the original death certificates and after the investigation are shown.

A special investigation on mortality in which the obtention of additional data is only in hospitals where death occurred proved to be, at least for under-fives, as regards causes of death, as good as any other investigation, with the advantage of being very much cheaper. The disadvantages include the impossibility of studying other factors such as family composition, feeding habits, type of medical attention received, time of residence at the place of death.

Special Investigation in Sudden and Unexpected Death in Adults

Sudden death in adults is becoming more important and several epidemiological studies have been undertaken in order to find out which are the risk factors and certain characteristics as regards occurrence, so as to ensure some preventive measures.

In Sao Paulo, knowledge as to frequency of sudden death in the general population by means of data registered in death certificates is almost impossible to obtain, because even when doctors do write down the proper cause of death and the sequency of causes up to the direct cause, very rarely do they put down the time elapsed between the beginning of the morbid process and death itself. Besides, hardly ever does one see the information "sudden death". Even when sudden death does occur in a person who was not receiving any medical attention whatsoever, for example on the street or in any other public place, and autopsy is performed, the death certificate indicates the findings but does not mention sudden death.

The knowledge and the study of sudden death in the population of Sao Paulo would only be possible by means of a special investigation, conducted as those that have already been done, in which the family would be interviewed in relation to the suddeness of the death.

During a period of twelve months (October 1, 1974 to Sept. 30, 1975), we undertook an investigation of mortality in adults, 15 to 74 years old, going along the same lines as in the "Inter-American Investigation of Mortality in Adults" (1962-1963)[3], save that data on variables that were not included before were now analysed according to multiple causes of death. A sample of one out of nine deaths in the above-mentioned age group of those residents in Sao Paulo was selected. Among the several analyses which are being performed the epidemiologic study of sudden death stands out.

The definition of sudden and unexpected death for this investigation was the same as that of Kuller et al.[6]: "An individual who died due to natural causes and who was not restricted to his home, a hospital or institution, or unable to function in the community for more than 24 hours prior to death, and in which the time interval from the onset of the fatal event until death was less than 24 hours".

Only some of the provisional results are being presented here as a more detailed analysis is underway. Such results are being shown as it is the first time that this sort of work has been undertaken amongst us.

The sample was of 2,738 deaths, of which 370 were due to non-natural causes (accidents, homicides, etc.) leaving 2,368 to natural causes. Of the latter 138 or 5.82% were due to sudden and unexpected deaths.

As regards the causes of sudden death, the following was observed:

	n.	%
Arteriosclerotic heart disease	55	39.85
Cerebrovascular disease	22	15.94
Dissecting aneurism of the Aorta	9	6.52
Hypertensive heart disease	9	6.52
Meningococcal meningitis	6	4.35
Other cardiovascular diseases	4	2.89
Other causes	5	3.62
Unknown causes	28	20.30
	138	100.00

The main cause of sudden death, as in studies performed in other countries, was the Arteriosclerotic Heart Disease, that in Sao Paulo was held responsible for 39.85% of the cases. In 20.30% of the cases we were not able to conclude towards a diagnosis, remaining this group as that of unknown causes; one should take note that in the 28 cases where it was not possible to conclude as to the cause of death in 15 (53.57%) the doctor had declared that death had occurred due to "Miocardial Infarction" but in fact with all the possible disposable information we were not able to come to this conclusion. The other causes of death are not very different from those which are observed in studies undergone in other countries. Nevertheless, Meningococcal Meningitis stands out as the 5th cause and that hasn't been observed in other studies. During this investigation in Sao Paulo an epidemic was underway.

As regards sex, a 2.18 ratio was observed as to male/female cases. As regards age, it was observed that the greatest percentage of sudden deaths occurred in the 35 - 44 year age group, in which 17.98% of all deaths were of the sudden type, a percentage that is by far superior to that of any other age group, as shown below:

Age	Deaths (natural causes)	Sudden Deaths n.	%
15-24	102	7	6.86
25-34	184	11	5.97
35-44	278	20	17.98
45-54	455	22	4.83
55-64	621	37	5.95
65-74	728	41	5.63
Total	2368	138	5.82

As to Arteriosclerotic Heart Disease (AHD), it was shown that 12.0% of deaths due to this cause were sudden ones. Some of the characteristics of the sudden deaths due to this cause were that the male/female ratio was 3.58 and as regard age, the greatest percentage of sudden deaths occurred in the 25-34 and 35-44 age groups, as follows:

Age	AHD-Deaths	AHD - Sudden Deaths n.	%
15-24	-	-	-
25-34	5	2	40.00
35-44	28	11	39.28
45-54	67	7	10.44
55-64	147	12	8.16
65-74	211	23	10.90
Total	458	55	12.00

As to duration of symptoms, that is the time elapsing between the first symptoms and death, the following was observed:

Duration	n.	%
Up to 15 minutes	20	36.36
15 min. to 2 hours	21	38.18
from 2 to 24 hours	14	25.46
	55	100.00

As to the habit of smoking, 50.90% of those who underwent AHD - sudden deaths were smokers. Arterial hypertension was present in 45.45% of the cases and diabetes in 21.81% of sudden deaths due to Arteriosclerotic Heart Diseases.

In 34.54% of the cases diagnosis was established by autopsy. In those cases where autopsy was not performed, the final clinical picture was suggestive of miocardial ischaemia there having been confirmation by electrocardiogram in the vast majority of cases where the interval between the initial symptoms and death was more than one hour.

In Appendix 5 data regarding death rates by all causes, natural causes, Arteriosclerotic Heart Diseases and sudden death rates are shown.

Other characteristics regarding sudden death, either by Arteriosclerotic Heart Disease or by other causes, such as temporal distribution, place of occurrence, type of occupation, marital status, preceding medical assistance and at the time of death, are being at present analysed and shall be presented in the near future.

As a general conclusion one can say that special investigations in mortality are extremely useful, allowing the correction of data, especially the causes of death. On the other hand, notwithstanding this aspect, they allow epidemiologic analyses regarding variables that are not usually part of the information registered in the death certificates.

References

1. Mac Mahon, B. & Pugh, T. F. - Epidemiology, Principles and Methods. Boston, Little, Brown, 1970.
2. Silveira, M. H. & Laurenti, R. - Os Eventos Vitais: Aspectos de seus Registros e Interrelaçao da Legislaçao Vigente com as Estatisticas de Saude. Revista de Saude Publica, Sao Paulo, 7:37-50, 1973.
3. Puffer, R. R. & Griffith, G. W. - Patterns of Urban Mortality. PAHO/WHO, Scientific Publication, No. 151, 1967.
4. Puffer, R. R. & Serrano, C. V. - Patterns of Mortality in Childhood. PAHO/WHO, Scientific Publication, No. 262.
5. Laurenti, R. - Causas Multiplas de Morte (Tese de Livre-Docencia Faculdade de Saude Publica, USP), Sao Paulo, 1973.
6. Kuller, L., Lilienfeld, A. & Fisher, R. - An Epidemiological Study of Sudden and Unexpected Deaths in Adults. Medicine 46(4) 341-361, 1967.

Appendix 1

Original Classification and Final Assignments to Some Causes of Death with Changes in S.Paulo, 1962/1963.(Inter-American Investigation of Mortality - Adults, 15 - 74 years)

Causes of Death	Original	Exclusions	Additions	FINAL
TOTAL	4361	1532	1532	4361
Infective and parasitic diseases	246	33	74	287
Malignant neoplasms	818	196	200	822
Cardiovascular diseases	1872	704	620	1788
Chronic rheumatic heart disease	89	19	36	106
Arteriosclerotic heart disease	499	80	122	541
Hypertensive heart disease	155	77	170	248
Diseases of the respiratory system	258	128	122	252
Influenza and pneumonia	126	57	23	92
Diseases of the digestive system	274	114	125	285
Ulcer of stomach and duodenum	29	2	19	46
Cirrhosis of liver (without alcoholism)	79	62	8	25
Cirrhosis of liver(with alcoolism)	28	6	63	85
Maternal Causes	24	3	15	36
Accidents and Violence	350	107	135	378
Remainder	519	246	238	510

Appendix 2

Infant Deaths from Certain Perinatal Causes as Underlying Causes Based On Death Certificates and on Final Assignments, with Ratios. São Paulo, 1968/1970(Inter - American Investigation of Mortality in Childhood)

Perinatal Causes		Original Certificate	Final	Ratio
TOTAL (760-778)*		1072	1119	1.04
Maternal conditions (760-763)		10	105	10.50
Difficult labor (764-768)		10	168	12.00
Other complications pregnancy, chilbirth(769)		33	247	7.48
Conditions of placenta,cord (770,771)		15	99	6.40
Birth injury, cause unspecified	(772)	58	52	0.90
Hemolytic diseases of newborn	(774,775)	20	26	1.30
Anoxic,hipoxic conditions	(776)	487	352	0.72
Immaturity	(777)	357	41	0.11
Other conditions of newborn	(778)	78	32	0.41

* ICD - 8th Revision

Appendix 3

Nutritional Deficiency as Associated Cause of Death in Children Under 5 years of Age (Excluding Neonatal Death) by Underlying Cause Group in S.Paulo Project and in 13 Latin American Projects Combined.1968/1970 (Inter - Americam Investigation of Mortality in Childhood)

Cause Group	S.Paulo			13 L.A.Project Combined		
	TOTAL DEATHS	With nutritional deficiency		TOTAL DEATHS	With nutritional deficiency	
		n.	%		n.	%
All Causes	2,354	1,108	47.1	21,951	10,349	47.1
Infective and parasitic diseases	1,191	710	59.6	12,598	7,667	60.9
Diarrheal diseases	844	529	62.7	8,770	5,331	60.8
Measles	156	74	47.4	2,103	1,311	62.3
Other	191	107	56.0	1,727	1,025	59.4
Nutritional deficiency	97			1,163		
Diseasesof respiratory system	525	181	34.5	4,469	1,435	32.1
Other causes	541	217	40.1	3,721	1,247	33.5

Appendix 4

Arteriosclerotic Heart Disease(AHD) associated with some other causes of death according to the original death certificates and the final results of the investigation. Sample of 1,832 hospital deaths,São Paulo, 1972/1973

Associated Causes	Original Certificates (AHD = 149)		Final (AHD = 200)	
	n.	%	n.	%
Infective and parasitic diseases	1	0,67	14	7.00
Malignant neoplasms	4	2.68	19	9.50
Diabetes	13	8.72	24	12.00
Hypertensive diseases	24	16.11	58	29.00
Cerebrovascular diseases	16	10.74	50	25.00
Diseases of arteries, arterioles and capilares	25	16.78	77	38.50

Appendix 5

Deaths rates by All Causes, Natural Causes, Arteriosclerotic Heart Disease and Sudden Death rates. Sample of 2738 deaths, 15 - 74 years, São Paulo, 1974/1975.

(Rates per 100.000 population)

Age	All Causes		Natural Causes		AHD		SUDDEN Deaths			
							Natural Causes		AHD	
	Number	Rate	Number	Rate	Number	Rate	Number	Rate	Number	Rate
15 - 24	191	132.91	102	70.98	-	-	7	4.87	-	-
25 - 34	272	232.02	84	71.65	5	4.26	11	9.38	2	1.70
35 - 44	355	390.03	278	305.43	28	30.76	20	21.97	11	12.08
45 - 54	499	783.54	455	714.45	67	105.20	22	34.54	7	10.99
55 - 64	665	1735.74	621	1645.64	147	389.54	37	98.04	12	31.79
65 - 74	756	3520.86	728	3390.46	211	982.67	41	190.94	23	107.11
15 - 74	2738	576.61	2368	498.69	458	96.45	138	29.06	55	11.58

MULTIVARIATE ANALYSIS OF HAND USAGE ON STRUCTURE AND FUNCTION

Dennis B. Gillings, Diane Makuc, Nortin M. Hadler
University of North Carolina at Chapel Hill

Summary

Three groups of female textile workers, each employed in different well-defined repetitive manual tasks for at least 20 years were identified in a single rural mill. Replicate data were obtained at one point in time for the following measurements on both hands: range of motion, degree of degenerative joint disease at each joint, malalignment at digital joints, osteophyte formation. Data items were either continuous or ordered categorical in nature and the joints of both hands for the same individual provided a multivariate profile of measurements.

Multivariable categorical data methods, and multivariate non-parametric and parametric techniques were employed to determine 1) observer agreement; 2) right and left hand differences; and 3) task differences. The statistical tests that were used are described.

It was concluded that there was 1) adequate observer agreement; 2) significant and consistent differences between the dominant right hand and the left; and 3) significant task related differences that were consistent with the pattern of usage in the industrial setting.

DESIGN OF STUDY

Selection and Description of Subject Groups. A single long established worsted mill in a small rural Virginia town was chosen. The mill employs over 600 people and is a major employer in a rural community with a stable population. It is characteristic of such plants that inter-job mobility is not commonplace. Three different manual tasks, burling, spinning, and winding that employ a large percentage of the workers were chosen because the task description has changed little over the past decades. Spinners and winders tend two different types of machines which (1) spin crude yarn into tighter thread and (2) wind several of the spun threads together for weaving. Burlers repair defects in the woven cloth.

Table 1 lists some characteristics of the study groups by task. Only female employees working continuously in the respective tasks for at least 20 years were considered eligible. Of those eligible, the number of volunteers is listed. No attempt was made to determine the motivations of those who elected not to participate.

Execution of the Clinical Study. The study was executed over the course of three working days. Subjects in groups of four were scheduled to arrive at 30 minute intervals at the on-site conference rooms made available by plant management. In a private room, they were individually interviewed by the research secretary and assigned a study number. The interview entailed completing a questionnaire and then reading and signing a consent form. During the remainder of the clinical exam, subjects were identified only by study number.

TABLE 1. DESCRIPTION OF STUDY GROUPS BY TASK

	Burl	Wind	Spin
Eligible Employees	39	16	20
Volunteers	32	16	19
Unavailable*	2	0	0
Excluded†	1	0	0
Total in Study	29	16	19

*Hospitalized with intercurrent illness.

†Inflammatory polyarthritis was detected in the course of examining this subject.

Clinical Examination. There were four examiners: three rheumatologists and a physical therapist. This team was repeatedly rehearsed prior to the actual study. The four subjects were randomly assigned to each of the four examiners. The examination required 15 minutes after which the same subjects were randomly assigned to a second examiner different from the first. Therefore, each subject in the study was examined by two separate investigators. The random assignment schedule was designed prior to the study.

Three categories of data were obtained by clinical examination:

1. The hands were examined for the presence of synovitis or the residua of major overt trauma. One subject (Table 1) was excluded from the study because of inflammatory synovitis and is currently under continuing management. Five subjects were detected who had incurred major trauma in the distant past with residual deformity to one digit.

2. The extremes of active range of motion were measured to the nearest 15 degrees with a small plastic universal goniometer. Fifteen degrees was chosen as the categorical unit of measurement as differences of this magnitude are less subject to doubts as to clinical significance. All small joints of the hand, the first carpometacarpal joint and the wrist were examined. An example of the data form used throughout the study is shown in Figure 1.

3. The circumferences of all distal (DIP) and proximal (PIP) interphalangeal joints were determined with an arthrocircameter.

Radiographic Examination. Radiographs were taken by the office technologist in the employ of an orthopedic surgeon in the community. Subjects were transported by the plant nurse in small groups in the weeks prior to the clinical study. A single postero-anterior radiograph of both hands and wrists was obtained. The radiographs were identified only by a numerical coded marker. Each

radiograph was evaluated by two rheumatologists without access to the code. Each joint for which goniometric data was obtained was scored for DJD from 0 to 4. In addition, the minimal width of the mid-phalangeal shaft was recorded to the nearest millimeter. Malalignment was determined by measuring, to the nearest 5°, the discrepancy in the alignment of the long axes of the contiguous bony shafts at the PIP and DIP joints, excluding the thumb.

Description of the Tasks. The task description was performed by Dr. M.A. Ayoub, a consulting industrial engineer and ergonomicist. A micromotion analysis was not performed. A standard time-motion analysis was made available by the plant industrial engineer. Dr. Ayoub supplemented this information with direct observation. Without access to the results of the study, he responded to a set of direct questions formulated by the investigators in order to rank the tasks by frequency of patterns of hand use.

STATISTICAL METHODS AND RESULTS

Description of the Groups of Workers and Their Tasks. The questionnaire was designed to provide some demographic data and to assess the homogeneity of the groups by task. The results are presented in Table 2. There were no significant task differences. Almost all subjects had lived their entire lives within the contiguous counties. Only two pairs of subjects were relatives. The groups did not differ in age (Table 2: Kruskal-Wallis test $P = 0.1769$). Further, there was no evidence of self-selection on the part of individuals as regards choosing which job to perform. Few persons admitted that they requested a position of spinning, burling, or winding (see Table 2).

It is apparent from Table 2 that the subjects perceived only winding as a bimanual task. All tasks are highly repetitive, stereotyped, and complex in that to some extent they are bimanual. However, in terms of frequency, the task analysis corroborates the perceptions of the subjects. Furthermore, winding differs distinctly from the other two because of a predominance of wrist motion and the employment of a power grip with little fine finger motion. Burling and spinning differ in that the latter tends to utilize a three-finger hand, sparing in use digits 4 and 5. These task descriptions were patently obvious results of the assessment undertaken. A more detailed description is not justified in the absence of an extensive micromotion analysis quantifying the force employed and frequency of use at each joint.

Data Analysis. Four separate response variables resulted from the clinical and radiographic studies: range of motion, malalignment, radiographic DJD score, and derivatized circumference. All data were obtained independently by two observers. The data analysis was planned to answer three main questions:

a) Were the observers in agreement as regards the data items recorded on each individual?
b) Were there differences between the right and left hands of each individual?
c) Were there differences between individuals who worked on different tasks?

The strategies that were determined to answer each of these questions are outlined below. Joint groups with trauma-induced deformity (five) and the two miscellaneous isolated missing values were excluded from the corresponding multivariate analyses. That two subjects considered their task as left-handed was disregarded except where indicated.

Analysis of Observer Agreement. The analysis of observer agreement was carried out to test agreement between observers as regards the measurement of individual subjects.

A descriptive assessment of observer agreement was obtained by computing the percentage of individuals with perfect agreement for both observers and the percentage of individuals with agreement for both observers within one unit of measurement. The degree of perfect agreement is presented in Table 3.

The measuring instruments were not really designed to have high perfect agreement, especially with respect to range of motion where measurements were taken within 15°. The percentages of individuals who show agreement within one unit of measurement are invariably high (data not shown), the majority being over 90%. Only 3 such percentages fell below 80%. The best agreement within one unit of measurement (100%) for range of motion is found in joints 9 through 12, and the worst, (71.4%) in joint 13 for the left hand.

Kappa-type statistics as described by Landis (1) provide a more refined measure of the extent to which two observers classify individual subjects into the same category. Weights are used to vary the definition of agreement. Kappa statistics were computed for each of the four response variables on joints with crude agreement less than 90%. Beyond 90%, agreement was felt to be more than adequate. For this analysis two sets of weights were considered, one corresponding to perfect agreement ($\hat{k}_1$) and the second to agreement within one measurement unit ($\hat{k}_2$). Kappa statistics can range from 0.00 to 1.00 with values from 0.00 to 0.20 roughly indicating "slight" agreement and 0.81 to 1.00 "almost perfect" agreement.

The data for right and left hands were considered independent in this analysis. All results were obtained with the computer program GENCAT (2) which is a general routine that analyzes categorical data by weighted least squares according to the formulation developed by Grizzle, Starmer, and Koch (3). The kappa statistics are presented in Table 4. The lowest value of $\hat{k}_1$ was 0.093 for range of motion at joint 7 and the highest was 0.538 for derivatized circumference at joint 13. All except three of the kappa statistics were significantly different from zero at the .05 level. For those measurements with agreement within one unit of measurement less than 90%, values of $\hat{k}_2$ ranged from 0.272 to 0.712. All of the kappa statistics for derivatized circumference and malalignment were significantly different from zero at the .01 level. The analysis using kappa

statistics supports the conclusion derived from the measures of crude reliability, namely that agreement between the first and second observers is acceptable.

Differences Between Right and Left Hands. Multivariate sign tests were applied where the right hand was ranked 1 if it had the greater score or measurement and 2 if it had the smaller. A midrank 1.5 was assigned if the right and left hands were equal. This test is based on the multivariate version of Friedman's test (4) as described by Gerig (5). The test compares mean ranks for each hand simultaneously across a profile of joints. Computations were carried out using the computer program FLOTA (2). The multivariate version of Friedman's test is equivalent to a multivariate sign test in the case where there are only two treatments (treatment corresponds here to right or left hand) in the same way that Friedman's test is equivalent to a (univariate) sign test when there are only two treatments.

The multivariate sign test was applied separately to each of the five digits and to the wrist. The results of the tests of significance are summarized schematically in Figure 2. It can be seen that there are extensive right and left hand differences. The multivariate test was significant in each case where data was available for range of motion and derivatized circumference. Several of the corresponding univariate tests were also significant, thus supporting these findings. None of the multivariate tests were significant for DJD score, although two of the univariate tests reached significance at the 0.05 level. There were no significant right and left hand differences in respect to malalignment.

Difference Between Tasks. Task differences were assessed in four different ways, referred to as Models 1-4.

Model 1 - right task hands only considered as the unit of observation.

Model 2 - left hands only considered as the unit of observation.

Model 3 - right and left hands considered separately, each hand being taken as an independent unit of observation.

Model 4 - right and left hands jointly considered as the unit of observation.

For each of Models 1-4 a multivariate Kruskal-Wallis test, as discussed by Koch (6) was applied to each profile of scores for the range of motion, malalignment and radiographic DJD scores. The means of the measurements for the two observers were the data points analyzed. The null hypothesis was that there were no differences among tasks. If the null hypothesis is rejected, then at least one of the tasks is different from the others. The direction of such differences may be clarified by inspection of the corresponding mean ranks. Joints were considered separately for univariate analysis and grouped by digit and wrist for multivariate analysis.

For this analysis, the treatments were taken to be the three tasks spinning, burling, and winding. The scores for each joint were combined across tasks and ranked. The multivariate Kruskal-Wallis test was used to compare tasks simultaneously across joints for each digit and the wrist using the computer program SPLOTA (2). Corresponding univariate Kruskal-Wallis tests (7) for each joint were routinely computed by SPLOTA to facilitate greater understanding of any differences detected. The results for this analysis and the MANOVA analysis described below are summarized in Figure 3.

The multivariate Kruskal-Wallis test was also appropriate for the derivatized circumference of joints. However, a MANOVA (8) analysis was selected as the procedure of choice as the data were continuous. Age was included as a covariate. The analysis was performed with task as the main effect. The null hypothesis was again tested for each of the models, right task hands, left hands, right and left hands separately, and both hands as the unit of observation. Hands with injured joints or missing data were eliminated from the analysis.

The most striking findings are summarized in Figure 3. Only p-values that were consistent in the several modes of testing including the task analysis, are presented in this figure. There is a problem of multiple comparisons with the large numbers of statistical tests and so consistency of results was chosen as the approach to handle this.

There were several task-related differences that were statistically significant for the left hand but not included in Figure 3 since these left-hand task-related differences were clinically unimportant resulting from differences within a single unit of measurement.

DISCUSSION

This study is cross-sectional and retrospective in design. It therefore suffers from the inherent flaw that we are unable to comment on the loss from the cohort that was first employed 20 years ago. Nonetheless we detect multiple and consistent differences in the structure and function of the hands of the women employed in the three tasks. It is our contention that these differences argue cogently for the rejection of the null hypothesis that pattern of usage does not influence structure and function of the hands. Attrition from the initial cohort is certainly multifactorial. However, attrition because of hand DJD would be expected to obscure the task related differences we observed.

Further, right-left differences were readily detected in the analysis of range of motion although such differences were demonstrable when derivatized circumference and DJD scores were analyzed (Figure 2). These demonstrations of greater impairment in the right hand by themselves argue for a role of usage (a traumatic element (9)) in the pathogenesis of primary DJD of the hands.

It is important to realize that this study is not designed to test for abnormality. We are testing the likelihood that a single independent

variable--the pattern of usage--influences the structure and function of the hands of three groups of women that are highly comparable. To test for abnormality one would need to identify a control group that was normal in terms of structure and function. The ideal control group would preferably lack the influence of the independent variable under study. Since that variable is hand usage, there is no ideal control. An alternative control group would be one in which all possible patterns of hand usage are represented without bias. If these patterns were defined and their frequency of occurrence in a suitable population known, then the usage patterns in a control group could be measured to demonstrate the absence of bias. However, it would be prohibitively expensive to identify a control group in this way. An alternative would be to take a simple random sample of middle aged women from a large population base. This would still have been potentially uninformative as no difference between experimental and control groups would not rule out the possibility that usage affects structure and function. Several types of usage would occur in the control group and some of these could affect structure and function more than others. One is then led to conduct a study in which the independent variable is manipulated in a more precise way. The above is an example of such a study.

REFERENCES AND NOTES

1. Landis, J.R. and G.G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159-174.

2. Development of GENCAT, FLOTA, SPLOTA computer programs, coordinated by Dr. Gary Koch, Dept. of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27514.

3. Grizzle, J.E., C.F. Starmer, and G.G. Koch. 1969. Analysis of categorical data by linear models. Biometrics 25:489-504.

4. Friedman, M. 1973. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. 33:675-699.

5. Gerig, T. 1969. A multivariate extension of Friedman's test. J. Am. Stat. Assoc. 64:1595-1608.

6. Koch, G.G. 1970. The use of non-parametric methods in the statistical analysis of a complex splitplot experiment. Biometrics 26: 105-128.

7. Kruskal, W.H. and W.A. Wallis. 1952. The use of ranks in one criterion variance analysis. J. Am. Stat. Assoc. 47:583-621.

8. Smith, H., R. Gnanadesikan, J.B. Hughes. 1962. Multivariate analysis of variance (MANOVA). Biometrics 18:22-41.

9. Acheson, R.M., Y.K. Chan, and A.R. Clemett. 1970. New Haven survey of joint diseases. XII: Distribution and symptoms of osteoarthrosis in the hands with reference to handedness. Ann. Rheum. Dis. 29:275-286.

TABLE 2

SPECIFIED CHARACTERISTICS BY TASK

Specified Characteristic	Burl	Wind	Spin	Total
Hand dominance				
right	28(96.6%)	15(93.8%)	17(89.5%)	60(93.8%)
left	1(3.4%)	1(6.3%)	2(10.5%)	4(6.3%)
Hand performing the task				
right	28(96.6%)	1(6.3%)	19(100.0%)	48(75.0%)
left	1(3.4%)	1(6.3%)	---	2(3.1%)
both	---	14(87.5%)	---	14(21.9%)
Task always done in the same way	29(100.0%)	6(37.5%)	18(94.7%)	53(82.8%)
Never lived outside county	18(62.1%)	11(68.8%)	14(73.7%)	43(67.2%)
A relative also worked at least 20 years burling, spinning, winding	1(3.4%)	3(18.8%)	2(10.5%)	6(9.4%)
Serious injury to hand				
right	3(10.3%)	1(6.3%)	3(15.8%)	7(10.9%)
left	2(6.9%)	3(18.8%)	1(5.3%)	6(9.4%)
At least five years at one other job	2(6.9%)	1(6.3%)	---	3(4.7%)
Have hobby with repeated manual work	5(17.2%)	1(6.3%)	---	6(9.4%)
Spend more than five hours/week at manual hobby	2(6.9%)	1(6.3%)	---	3(4.7%)
Spent more than five years at manual hobby	3(10.3%)	---	---	3(4.7%)
Currently have manual hobby	3(10.3%)	1(6.3%)	---	4(6.3%)
Requested position in burling, winding, or spinning	4(13.8%)	1(6.3%)	1(5.3%)	6(9.4%)
Number of cases	29(100.0%)	16(100.0%)	19(100.0%)	64(100.0%)
Age (mean ± S.D.)	56.2 ± 7.7	49.0 ± 6.1	49.4 ± 6.0	

TABLE 3

PERCENT OF SUBJECTS WITH PERFECT AGREEMENT BETWEEN OBSERVATIONS

	Range of Motion		Derivatized Circumference		Malalignment		DJD Score	
Joint*	Right	Left	Right	Left	Right	Left	Right	Left
1	56.3	60.9					96.9	92.2
2	38.7	47.6					79.7	89.1
3	37.5	34.4					75.0	73.4
4	38.1	53.1					92.1	92.2
5	46.9	48.4					73.0	78.1
6	54.7	53.1					65.6	79.7
7	50.0	40.6					87.5	89.1
8	42.2	50.0					62.5	75.0
9	82.5	87.3	71.4	60.3	76.2	76.2	68.3	50.8
10	93.8	93.8	71.9	81.3	68.8	79.7	67.2	67.2
11	93.8	87.5	76.6	89.1	59.4	65.6	68.8	71.9
12	88.9	90.6	61.9	75.0	71.4	64.1	69.8	65.6
13	34.4	36.5	73.4	84.1			50.0	61.9
14	73.0	75.8	74.6	66.1	60.3	54.1	49.2	59.0
15	76.6	74.6	75.0	84.4	84.4	57.1	54.7	74.6
16	79.7	82.8	76.6	87.5	64.1	45.3	46.9	56.3
17	76.2	79.7	63.5	73.4	68.3	59.4	49.2	67.2
Number of hands	64	64	64	64	64	64	64	64

*Numbering of joints is explained in Figure 1.

TABLE 4

KAPPA STATISTICS* FOR RANGE OF MOTION, DERIVATIZED CIRCUMFERENCE, MALALIGNMENT, AND DEGENERATIVE JOINT DISEASE SCORE BY JOINT

	Range of Motion		Derivatized Circumference		Malalignment		DJD Score	
Joint	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_1$	$\hat{k}_2$
1	0.322	NA					NA	NA
2	0.195	0.418					DEGEN§	NA
3	0.202	NA					0.274	NA
4	0.168	0.272NS					NA	NA
5	0.151	NA					DEGEN§	NA
6	0.210	NA					0.125	NA
7	0.093NS	NA					DEGEN§	NA
8	0.253	0.607					0.128NS	NA
9	0.386	NA	0.436	NA	0.459	NA	0.226	0.399
10	NA	NA	0.472	NA	0.137	NA	0.266	NA
11	NA	NA	0.465	NA	0.324	NA	0.337	NA
12	NA	NA	0.346	NA	0.375	NA	0.174	NA
13	0.225	0.332	0.538	NA	---	---	0.216	0.447
14	0.414	NA	0.516	NA	0.292	0.712	0.267	NA
15	0.345	NA	0.468	NA	0.371	NA	0.250	NA
16	0.470	NA	0.531	NA	0.321	NA	0.282	NA
17	0.320	NA	0.457	NA	0.264	NA	0.370	0.590
Number of hands	128	128	128	128	128	128	128	128

*$\hat{k}_1$ measure perfect agreement and $\hat{k}_2$ agreement within one unit of measurement.

All values are significant ($p < .05$) except where indicated by NS.

NA = not applicable because crude agreement >90%.

§DEGEN = not appropriate because of degenerate contingency table.

RIGHT HAND

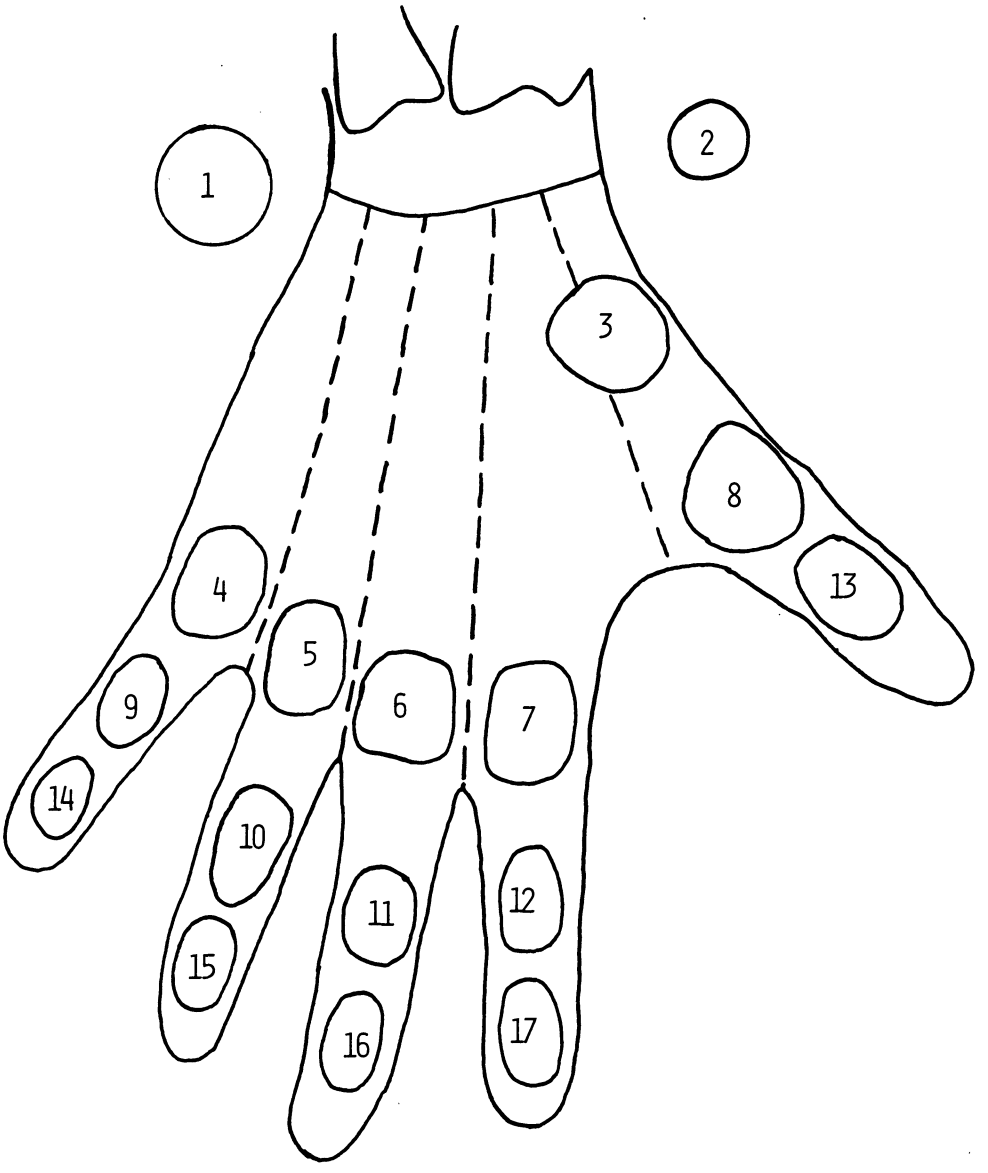
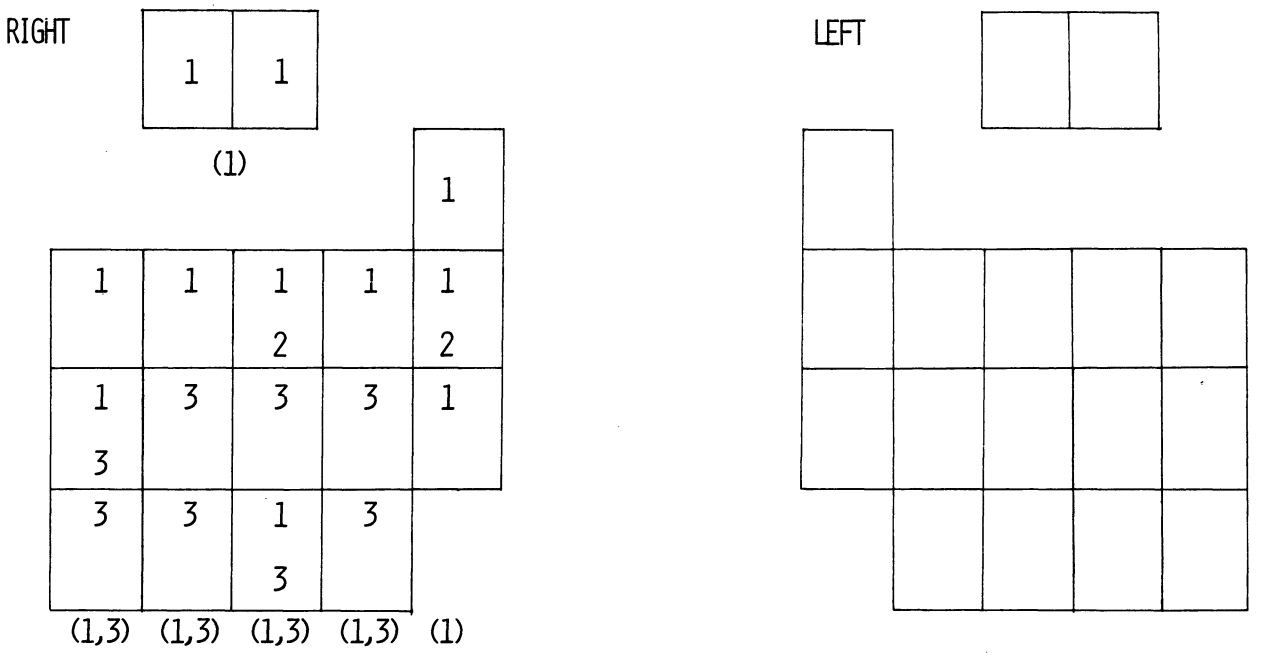


Figure 1

Data collection form and joint numbering code for the right hand. The mirror-image form was used for the left hand. For range of motion data, joint 1 represented the range of deviation at the wrist and 2 represented the range of extension-flexion. For radiographic DJD score joint 1 was the radial-carpal joint, 2 was the inferior radial-ulna joint. Joint 3 represented the first carpo-metacarpal joint throughout. Joints 4-17 are the anatomically corresponding digital joints.



1. Range of Motion
2. D.J.D. Score
3. Derivatized Circumference
4. Malalignment

Figure 2

Differences between right and left hands. The format of this figure is a stylized version of Figure 1. Rejection of the null hypothesis utilizing the multivariate extension of Friedman's chi-square test is indicated by a symbol designating the more impaired side. If the entire digit or wrist is significantly impaired, that is indicated by the designation in parentheses beneath the corresponding region. The data from which the null hypothesis is rejected is indicated numerically: 1 for range of motion, 2 for radiographic DJD score, 3 for derivatized circumference and 4 for radiographic malalignment.

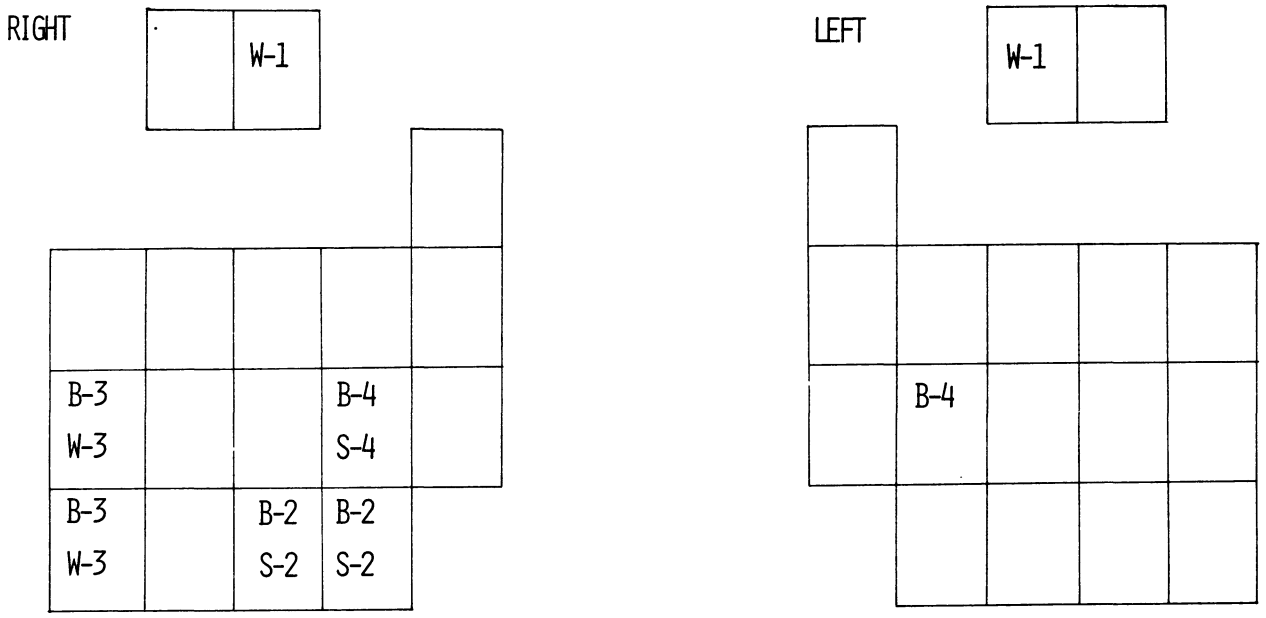


Figure 3

Task-related differences. The format of this figure is a stylized version of Figure 1. The most consistent and statistically significant structural and functional impairments deduced from the Kruskal-Wallis and MANOVA analyses are noted. When the null hypothesis is rejected, the tasks are ranked and the task(s) with the most impairment is indicated in the box corresponding to the involved anatomical region. The notation for each task is W for winding, B for burling, and S for spinning. The data from which the null hypothesis is rejected is indicated numerically: 1 for range of motion, 2 for radiographic DJD score, 3 for derivatized circumference and 4 for radiographic malalignment.

DISCUSSION

Adela Legarreta, Universidad de Chile

Although the concepts of Epidemiology and Health Statistics are surely familiar to most of you, I have considered it convenient to start these few comments by referring to them, to assure a better understanding of what I shall say thereafter.

1. What is the present concept of Epidemiology?

Historically, Epidemiology had its origin in the study of the great epidemics of the past (cholera, the plague, smallpox, etc.). The methods which were then developed were rapidly applied to all infectiuos diseases, without considering whether or not they were epidemic.

In developed countries other types of disease have gained importance and constitute the major health risks which originate, to a large extent, in the great environmental changes and modes of living imposed by industrialization and the consequent migration of the population to cities, as well as the considerable numerical and proportional increase of population groups of a more advanced age. Among these diseases which now prevail in many countries are cancer, hipertension, the coronary diseases, diabetes, peptic ulcer, mental disorders, etc. Kerr L. White has pointed out recently that the application of the concepts, methods and principles of Epidemiology to the fields of administration and research of health services constitutes an extension of their application to other problems of health and diseases.

Within the present epidemiological landscape, a group of phenomena stand out on account of their magnitude, their transcendence and their diffusibility. Among them are violence, delinquency, rape, crime, and the phenomena of evasion such as alcoholism, drug addiction and, suicide.

There exist numerous definitions of Epidemiology. For some it is the study of the states of health in human populations, as its area of interest is not confined to disease, but comprises also other biological processes such as growth, multiple pregnancy, fertility, etc.

The broadest definition which we have found is the following: "It is the scientific method applied to the study of the health of a human group".

2. Health Statistics

An international agreement has not yet been reached for defining Health Statistics. In their broadest sense, one could say, that consists of the application of the statistical method which provides the techniques for collection, elaboration and analysis of information relating to the health of a population. This information refers to the state of health of populations, to the conditioning factors of that state of health, such as physical, environmental and social conditions, and to the resources and activities of health services.

A series of generally organized systems on a national scale, allow for the qualifcation of a series of aspects which reflect on health, is considered within this type of statistics. Some systems which, even though they pursue manifold ends, are employed for health objectives, are so included in Health Statistics. A non-exhaustive list of the systems which Health Statistics comprises, are the following:

2.1 Population statistics obtained from census figures.

2.2 Vital statistics obtained from the Civil Registration services, specially dealths, fetal deaths and births.

2.3 Morbidity statistics.

2.4 Health resources statistics.

2.5 Health care statistics.

2.6 Statistics which refer to the physical, cultural, social, economic, etc., environment.

It can be generally stated that all statistical information refers to a community can be considered, in a broad sense, as falling within the realm of Health Statistics, when this information is used for the purpose of finding out about health characteristics of the population, and the measures adopted for health promotion, protection, restoration and rehabilitation.

So far we have referred to Health Statistics as being a set of statistical systems; this is, however, only one of their aspects. A second aspect is the application of the statistical method to specific problems in the health field. This is the investigation in Health Statistics.

3. The Application of Health Statistics to Epidemiology

Here we must distinguish both aspects of Health Statistics:

3.1 Use of data supplied by statistical systems. The information resulting from these systems has numereous applications such as the analysis of behaviour in time of mortality, morbidity, fertility, etc. Study of the quantity and type of health services available to the population, as well as the efficiency of specific medical care resources. These and other problems belonging to the scope of Epidemiology may be studied by taking advantage of the information which has been gathered and processed in the different systems which make up Health Statistics. They have the advantage of national coverage of this type of information in most countries, of the fact that collection and processing are normalized, and that they are of a continuous character. They are used mainly in epidemiologic studies of a descriptive type. Their major drawbacks are the limitation of the

data which may be analyzed, which is precisely determined in each system, and the not always satisfactory quality of information gathered.

Epidemiology uses not only the already processed information, but also the basic documents of these systems, a procedure which makes for greater flexibility. There exist numerous research studies which use as their data source medical death certificates which permit the analysis of mortality characteristics (studies of causes of death according to sex, age, geographical distribution, socio-economic level, etc.). Other basic documents of health systems which are used for epidemiological purposes are statistical reports of hospital patient discharges, which give information on hospital morbidity, as well as daily registration forms of medical consultations in clinics of health institutions, which contain data on ambulatory morbidity.

3.2 Research in Health Statistics

We have defined research in Health Statistics as the application of the statistical method to the study of specific health problems. This is a controversial field, as the statistical method is only an instrument of the scientific method; it is, therefore, this last method which is really used for research in the health of populations. This is an ambiguous ground on which Epidemiology and Health Statistics become confused. The discussion which arises on whether Health Statistics include Epidemiology, or vice-versa, is not strange. In fact, many of the pioneers in medical statistics, such as FARR, GREENWOOD and PEARL, dealt with problems which are now classed as epidemiological.

Health Statistics and Epidemiology superpose each other in their areas of application. Epidemiology, as the study of the health of a population, makes use of many tools, one of which is statistics. Statistics, as a discipline which is applied to observations of groups of individuals, has many uses in different branches of science, one of which is Epidemiology.

In short, both types of disciplines coincide in their objective, which is the study of health, in the employment of the scientific method, and in the scope of study, the community. The same methods of statistical analysis are also used (differnt types of regression, life tables, non-parametric methods, parametric and non-parametric analysis of variance, etc.).

The discussion about whether the study of the health of populations is a question to be dealth with by Statistics or Epidemiology is only a matter of academic importance, what seems to be really important, is the collaborative work among those who cultivate both of these disciplines. This not only redounds upon the technical quality of the investigations undertaken, but also upon a greater knowledge and experience for both types of of professionals.

DISCUSSION

Erica Taucher, Latin American Demographic Center (CELADE)

The papers that we are discussing today are interesting examples of the variety of approaches and meanings that can be given to epidemiological studies. First, let us refer to Lilienfeld's historical review. Not only in epidemiology but in most applied sciences, the search for analytical methods and the extensive possibilities of computer techniques lead us often to forget a basic question: where did we come from and where do we go from here? Under these circumstances it is useful to remember that scientific reasoning is not new and that it cannot be replaced by methods or computers. Although basically I agree with the final statement of the authors, that men and methods make statistics, I think that another element has been omitted in their paper, that is: the problems that should be solved after having been analyzed through these methods by these men. Those problems are essentially the ultimate objective of any epidemiological study. The kind of problems that are perceived in different times or in different stages of socioeconomic development call for different approaches or analytical methods--which explains part of the variety mentioned earlier. The other two papers are good examples exemplifying these comments.

Through his report on some special investigations in mortality, Laurenti emphasises the importance of two problems which interfere with epidemiological studies: the availability and the quality of data. In many countries, mortality statistics are still one of the only possible data sources for epidemiological studies, although everyone agrees that they are by no means the ideal health indicators. Moreover, studies like those described by Dr. Laurenti are not even possible in a great number of developing countries where vital statistics are often very unreliable. This paper therefore deals with a situation that we could describe as being of intermediate or high development. In consequence the methods to test the quality of data or to complement them could be considered useful and could be transferred to the analysis of other problems only in countries or areas with a similar availability of mortality data.

Finally, in Gilling's paper we find an example of the possibilities of health research when resources are available to collect the needed data. Under these conditions, after the design of the research project, the analytical methods shift to a more important place. In this special case they required the support of computer facilities. Although I would prefer to remain on a general level in my comments, I would like to express a doubt regarding the test of agreement between observers. I am wondering if it would not have been better to explore whether the differences between hands or between tasks were consistent for different observers on the same individual. Since the differences were the main problem, this kind of agreement might have been of more interest than that of the absolute value of the measurements.

I would like to go back to the more general view. In summary, what we can conclude from the different papers is that if we accept that epidemiology is aimed at discovering and quantifying health problems, the stress on the collection of data, on their quality, or on the development of new analytical methods depends largely on the particular circumstances under which the research is performed.

THE SOCIAL EXPERIMENTS: STATISTICAL ISSUES AND THE ROLE OF STATISTICIANS

Charles E. Metcalf, Mathematica Policy Research [1]

A. INTRODUCTION

In the past ten years the methodology of controlled experimentation has taken firm hold as a focal point of analyses of major changes in social programs. Beginning with the New Jersey negative income tax experiment, there have been several major controlled experiments--with randomized assignment of participants to treatment programs and a control group--in the areas of income maintenance, housing allowances, health insurance, and employment programs. Furthermore, the existence of such experiments has had a major impact on the methodologies used for program evaluations not employing randomly selected control groups.

A major apparent concern to social scientists outside the economics profession is not so much the experiments themselves, but their own lack of involvement in the design and execution of the experiments. Critiques of the New Jersey experiment by sociologists have stressed how the analysis would have been improved by the inclusion of more (or higher quality) sociologists. The American Statistical Association has now scheduled a session on "the role of the statistician in social experimentation" and further compounded the issue by inviting an economist to deliver a paper on the topic.

While a broad range of the social sciences was represented in the major social experiments--sociologists, psychologists, political scientists, and statisticians--it is true that economists have played a dominant role in the recent growth of social experimentation. With regard to issues of sampling and statistical methodology, some statisticians were utilized, but there was a clear tendency for economists to call upon econometricians with training in statistics rather than upon individuals regarding themselves as statisticians.

These developments in the social experiments were an extension of tendencies already apparent in economics as a profession. Economists have been more heavily involved in quantitative measurment and hypothesis testing than have sociologists, and in the process econometrics became a sophisticated "in house" form of applied statistics. The earlier income maintenance experiments in particular were primarily intended to test hypotheses about economic behavior already well developed in economic theory--but with a data set not subject to the limitations inherent in the historical data commonly used. Despite the methodological departure into the realm of controlled experiments, the social experiments have retained both the perspective and the analytic approach of economists.

The sociologists' lament concerned economists' perspective about the objectives of the experiments--i.e., whether the right questions were being asked. The absence of participation by statisticians, if in fact true, relates to whether executers of the experiments availed themselves of available expertise when dealing with the statistical problems inherent in social experimentation.

There are four major areas where scientific expertise generally belonging in statistics as a disciplinary classification was required for execution of the social experiments: (1) establishment of the sampling frame, (2) experimental design, (3) empirical estimation and hypothesis testing, and (4) extrapolation of empirical results to quantitative measures relevant for policy decisions. While these four areas are closely interrelated I would like to discuss each area in turn, stressing the statistical issues involved, and speculating about what the impact of greater involvement of statisticans might have been. I shall begin by specifying a prototype design model with a simplified structure--having clear statistical implications but corresponding to none of the experiments actually performed. With specific examples from the New Jersey Experiment and from the other major experiments, each deviation from the prototype model can then be judged both in terms of the advantages claimed to justify it and the statistical problems associated with it.

The remainder of this introduction summarizes the key features of some of the major social experiments to date and outlines the prototype design model. Sections B through E address the four areas of scientific expertise cited above, while section F provides some concluding remarks.

The Major Social Experiments

The following research projects all utilized some form of randomized assignment of individuals or households into a treatment group, who were eligible for participation in the social program being evaluated, and a control group who were not eligible for program participation but whose behavior was observed. The list is not meant to be exhaustive; rather it includes the major income maintenance experiments plus illustrative social experiments in other areas with which I have some familiarity.

The New Jersey Income-Maintenance Experiment, funded by the Office of Economic Opportunity beginning in 1967, was the first and most widely scrutinized of the social experiments.[2] Several forces joined to bring about the experiment--increasing advocacy of negative income taxation as a viable policy option, combined with continuing concerns about the potential effect of universal income maintenance on work incentives; the feeling in both the scientific and political communities that evidence on work incentive effects obtained to date with non-experimental

techniques was insufficient; rising interest in social experimentation as a viable research option; and a proposal to undertake such an experiment by Heather Ross, then a Ph.D. candidate in economics at M.I.T.[3] The primary research question to be addressed was the effect of a negative income tax (N.I.T.) on the labor supply of families with a prime age, able-bodied male. A sample of 1,357 such families was drawn from five urban centers in New Jersey and Pennsylvania,[4] and 725 were allocated to a variety of NIT plans for a period of three years. Including design, staggered payment periods, and subsequent analysis, the experiment lasted seven years and cost $7.6 million dollars.

The Rural Income Maintenance Experiment applied a similar experimental design to 800 families in rural areas of Iowa and North Carolina. Its smaller sample size was fragmented by the designation of 100 sample points as households with a female head, and 100 as households with a head over 55 years of age; and by the stratification by farm/non-farm status as well as by geographic location.

The Gary Income Maintenance Project utilized a predominantly black sample of 1,600 families in Gary, Indiana, over half of which were families with female heads already receiving Aid for Dependent Children.

The Seattle/Denver Income Maintenance Experiment is the last of the U.S. income maintenance experiments and the most ambitious. It has a larger sample (5200); varies the duration of payments (3, 5, and 20 years) to test long-run program effects; includes job counseling and training for a fraction of the sample; and utilizes a non-linear tax rate structure for some treatment plans.

A flurry of interest in experimentation in Canada almost led to a similar array of Canadian experiments, but only the Manitoba Minimum Annual Income Project got off the ground, with a New Jersey style experimental design in Winnipeg and a "saturation" site in Dauphin, Manitoba.

Outside the realm of income maintenance, a series of three Housing Allowance experiments made payments to low income households either conditional on meeting minimum housing standards, or determined as a percentage subsidy of rental expenditures. The Health Insurance Study involved randomized assignment of individuals to different health insurance plans.

Supported Work provides transitional employment to ex-addicts, ex-offenders, AFDC mothers, and youth groups believed to have special problems adapting to conventional jobs. Objectives of the program include measurement of post-program employment, criminal behavior, drug use, and within-program productivity. A simple random assignment to program and control group status is made among applicants meeting eligibility standards. Colorado Monthly Reporting examines the effects of converting a random sample of Denver AFDC recipients to a computerized payment system based on the previous month's income, rather than the conventional method of a needs determination every six months; in addition, the entire caseload of Boulder County was placed on a monthly reporting system.

The Prototype Model

To facilitate the evaluation of the four critical areas where decisions relating to statistical methodology were made in the social experiments, the following prototype model can be a useful point of departure. The prototype model involves four explicit steps:

(1) The drawing of a simple random sample of all households in the U.S.;

(2) Simple random assignment of a program treatment to half the sample, with the remaining sample designated as the control group;

(3) Direct comparison of treatment and control groups to measure program effects, with simple difference-of-means statistical tests; and

(4) The interpretation of results as measuring directly what would happen with full scale adoption of the program.

While the above model is appealing in its simplicity, there were felt to be persuasive reasons for making major departures from the prototype at all four steps.

B. THE SAMPLING FRAME

None of the major experiments utilized a simple random sample of the sort envisioned in the prototype model. Major issues debated by designers of the experiments included (1) restriction of the sampling universe to policy-relevant subsets of the population; (2) the adoption of dispersed sampling versus implementation of "saturation" experiments; (3) the use of a random national sample vs. "test bores" of the population in a small number of sites; and (4) sampling procedures within sites.

The first major departure from the prototype model involved the truncation of the sample universe to include only families meeting certain income-eligibility and demographic criteria. Most of the programs being considered were targeted to particular segments of the population. For example, the NIT is targeted to low income households, even though all members of the population are eligible if their incomes fall into the relevant range. Thus the decision was made to sample only low income households rather than to observe large numbers of higher income people who were unlikely to receive program benefits. This truncation of the sample by income level was the most severe in the case of the New Jersey experiment, which included only families with incomes below 150% of a commonly used poverty income definition. The sampling universe for the New Jersey experiment was further restricted to include only families with a prime-age, able-bodied male. Since the labor force response of

such families was believed to be pivotal in the evaluation of a universal income maintenance scheme, it was decided to concentrate efforts on testing hypotheses about one group rather than to spread resources across heterogeneous family types.

Such sample truncation led to at least two problems. First of all, since income is under the partial control of household members through labor market decisions, the truncation variable was closely related to behavioral responses being measured in the experiment. It has been well established that when a sample truncated to restrict the domain of a variable is used to estimate determinants of that variable, conventional regression techniques will lead to biased results.[5] Secondly, it was realized ex-post that the truncation process eliminated the possibility of measuring program effects of major interest. Specifically, there were many two earner families who would receive NIT payments if one of the earners quit his or her job, but who had incomes in excess of 150% of the poverty level so long as both jobs were retained. Thus the severe truncation of the New Jersey experiment prevented a proper test of this response by excluding such families from the sample; available evidence suggests that this would be one of the largest sources of work reduction as a result of the NIT.

Subsequent experiments alleviated the problem by truncating the sample at a much higher level of income, and by including a broader range of demographic groups. At least one experiment went too far in the direction of sample heterogeniety. With a sample of only 800, the Rural experiment included female and aged household heads as well as families with prime-age males and stratified by geographic region and farm/non-farm status. The result was a sample with very little power for testing behavioral hypotheses.

The prototype model implies the choice of a dispersed sample rather than a saturation sample. One might expect that an individual's response to a program in which he participates as part of a random sample may differ from one where all individuals like him in the same community are subject to the same program. Also, saturation may be required to measure community effects on non-program participants; to observe major economic responses to the program (e.g., the effect on housing supply of a major housing subsidy program); and to evaluate the operational feasibility of implementing a full scale program. The idea of a saturation experiment including all program-eligibles in a random sample of localities has been frequently discussed but never implemented. The rejection has usually been on cost grounds. Several of the social experiments did, however, include saturation of selected sites without the explicit use of corresponding control groups--in particular, the Housing Allowance Supply Experiment and portions of the Manitoba and Colorado Monthly Reporting experiments.

The rejection of a random national sample in favor of a "test bore" sample in a geographically limited area was one of the more controversial decisions, made first in New Jersey and replicated in subsequent social experiments. Against the obvious loss of all statistical power for national extrapolations were placed the following advantages of a "test bore" sample: the (ultimately dominating) issues of cost and administrative feasibility, and the ability to test hypotheses against a homogeneous background environment. Given the limited resources available for experimentation, increasing the power of within-sample hypothesis testing was felt to be more important than representativeness of the sample. Most people involved with the experiments continue to believe that this decision was a correct one, at least given the information available at the time. I suspect that if statisticians rather than econometricians and policy analysts had led the movement to social experimentation, however, the case for a national sample would have been more forcefully defended.

The lack of a statistician's orientation also had an impact on the sampling procedures within sites, particularly in the New Jersey experiment. Because the yield of low-income households from New Jersey screening interviews was much lower than anticipated, cost considerations required that enumeration of the sampling frame be limited to census tracts with a high incidence of low income households. Thus poor families in low density areas had a zero probability of selection, and this fact resulted in the major unanticipated feature of the New Jersey sampling frame: the predominance of blacks and Puerto Ricans, and the resulting foray to Scranton, Pennsylvania in search of poor whites.

C. EXPERIMENTAL DESIGN[6]

In our prototype experiment, we observed the effects of a single policy change and interpreted the results directly. Some of the social experiments--Supported Work, for example--adopted designs similar to the prototype. The New Jersey experiment and most of its successors deviated from the prototype, however, both by including a multiplicity of experimental treatments and by adopting a complicated method of assigning sample households to specific treatments.

The case for a more complex experimental design rests on three major arguments:

(a) Policy interest is focused not on a single, known program but rather on a range of programs with similar characteristics.

(b) The experimental environment cannot provide a direct test of the relevant policy issues; thus the experimental design must provide the necessary information for extrapolating the results to the appropriate environment.

(c) An efficient experimental design should reflect prior knowledge about the structure of hypotheses to be tested and about differential costs of alternative experimental treatments.

From the nature of the reasons given, it should be fairly apparent that decisions regarding experimental design, hypothesis testing, and extrapolation of results are closely intertwined. Despite this, I shall maintain the fiction of distinguishing among the three areas, and outline the principles of experimental design developed for the New Jersey experiment.

Let us first consider a "design space" of potential program treatments as a range of testable programs of direct policy interest. If, for instance, we knew that our policy choice were limited to a single negative income tax plan versus no plan at all, we might limit our design space to a single plan plus a control group, as in the prototype design. If our objective were to choose among three plans, we might opt for a design space with three corresponding plans in addition to a control group. Such a design would permit a comparison of behavioral responses between any two treatments, and between a single treatment and the control group.

An increase in the number of treatments given a fixed budget or total sample size obviously reduces the number of observations per cell, and thus the precision of any pairwise test. It is desirable to develop some method of exploiting similarities among responses to alternative treatments not only to alleviate the loss of precision involved in testing multiple treatments, but also to make statements about behavioral responses to similar treatment plans not explicitly included in the experiment. This latter issue can be of major importance if the set of policies having potential policy interest shifts during the course of the experiment.

The notion of similarities among treatments suggests an alternative view of the design space as a range of program characteristics that affect household behavior, with a range of plan characteristics rather than merely specific treatments being of direct policy interest. In the case of the New Jersey experiment, each NIT plan was defined by a specific combination of income guarantee, G, and tax rate, t. The motivation behind restating each treatment in terms of characteristics influencing behavior came from the added assumption that behavioral responses vary in some continuous manner with variation in plan characteristics. If the relationship of behavior to variations in G and t can be approximated by a continuous response function of known (maximum) dimension, a design space including values for G and t at the extremes of the range of potential policy interest, plus sufficient interior values to identify the assumed response function, provides information about a complete continuum of policy options rather than simply a limited set of specifically tested alternatives. Correspondingly, precision in the estimated response at a specific G and t combination is derived not only from observations at that point, but from all observations relevant for identifying the "response surface." Because extrapolation of the effects of G and t combinations beyond the extreme observed variations can be done (if at all) with less confidence, the emphasis in this framework shifts to specifying the extremes of potential policy interest rather than the points of greatest direct policy interest. That is, even if we are most interested in obtaining information about central regions of our "policy space," this interest may be best served in an experiment stressing treatments at the fringes of our range of interest.[7/]

Once we think of obtaining information about a policy space in terms of testing hypotheses relating to household responses to program characteristics, plans to be included in the design space and those of direct policy interest may no longer coincide. Even if we know with certainty the policy alternatives to be considered, the optimal experimental design could, under some circumstances, not only exclude certain treatments of direct policy interest, but also include other treatments not among the set of policy alternatives.

The range of treatment plans can also be defined in terms of characteristics required to extrapolate results to a nonexperimental setting (see Section E). For example, the income maintenance experiments were of limited duration, whereas the programs of policy interest are presumably permanent. In order to permit the appropriate projections to be made, the Seattle/Denver experiment systematically varied the duration of its treatment programs from a minimum of three years to a maximum of 20 years.

The above discussion indicates that a careful consideration of experimental objectives could result in a design space that differs from a simplistic statement of programs of direct policy interest. Similarly, a sample allocation process that takes into explicit account both specific experimental objectives and budgetary and other constraints may lead to a violation of some commonly proposed principles relating to orthogonality of the sample design--these principles involve relationships (1) between plan assignments and household attributes and (2) among plan characteristics or variables to be included in an estimated behavioral model.

Given a situation where the design space is defined as a set of policy alternatives and where a decision has been made regarding the number of households to be assigned to each plan, it is often proposed that households for each cell be chosen by a self-weighting random sampling procedure. Even if the aggregate sample is to be stratified by certain household attributes, the view holds that the stratification characteristics should not influence the probability of assignment to a specific plan. That is, orthogonality of plan and stratification characteristics would be maintained so that simple comparisons could be made across plans.

Orthogonality is also typically stressed as a desirable feature of sample allocations across design spaces dimensioned in terms of plan characteristics because it permit hypotheses concerning a single characteristic to be tested without having to control for variations in

other plan and stratification characteristics. Indeed, orthogonality is an optimality condition for a class of problems often discussed in the design literature.[8]

Consider a case where the objective of the experiment has been defined as measuring experimental response relative to the control group for each of several treatments. In this case a regression form of an analysis of variance framework suggests itself where household behavior is viewed as a linear function of a set of dummy variables (one for each plan), and where the goal is to obtain accurate estimates of the coefficients associated with the differential effect of the experiment at each design point. If we specify the objective to be the minimization of a weighted sum of coefficient variances given a budgetary restriction, the optimal allocation of households could correspond to the uniform distribution across plans proposed above--*if* equal weight is attached to each variance term *and* costs per observation are identical across plans.

This latter condition is violated in the case at hand, since an intrinsic feature of NIT plans is that costs per observation vary systematically with plan characteristics--namely, the guarantee level and the tax rate. Starting from an initial uniform allocation where sample points of differing costs make the *same* marginal contribution to the experimental objective, the efficiency of the design could be improved by surrendering some expensive observations for a larger number of cheaper ones.

This latter result strongly influences the allocation of households to the control group; which is far less expensive per observation than the experimental cells. For instance, in order to measure with minimum variance the differential behavior between a control group and a single experimental cell in a situation where the cost per observation for experimentals is nine times that of controls, 75 percent of the sample should be assigned to the control groups and only 25 percent to the experimental treatment. Compared with an allocation of 300 treatment observations and 900 control observations, moving to equal cell sizes (360 each) would increase the variance of our estimate by 25 percent. Given the cost assumption which generated the three-to-one ratio between controls and experimentals, 75 percent of the budget is still expended on experimental plans. Thus, other things equal, a more expensive plan would be allocated a smaller number of observations in the optimal design, but would command a larger share of the experimental budget.

Cost differentials also play a role in the decision whether or not to stratify the sample by household characteristics. (The other major consideation is whether identification of differential responses by household characteristics plays an explicit role in the experimental objective.) If the population of interest were stratified by characteristics which affected costs per observation (e.g., family size or income), and if the experimental objective were to estimate the mean population response to a given treatment, the optimal strategy would be to oversample in those subgroups for which information could be obtained more cheaply.

It should be apparent that accounting for cost differentials in the sample allocation process is sufficient to destroy orthogonality between experimental variables and population characteristics as an optimality condition. Because the cost differential between experimental and control observations depends on household income, for example, the probability of assignment to a particular cell would no longer be independent of income. Basic principles of randomization are retained, however, if all households within a single stratum (that is, households identical in terms of stratification characteristics) face the same set of assignment probabilities.

The sample allocation models used in the income maintenance experiments went further than simply to account for variations in observation costs. The Conlisk-Watts model[9] which formed the basis of sample allocations in the income maintenance experiments has four major components:

(1) an assumed structural relationship, specified as a regression model, relating behavioral responses to treatment and household characteristics;

(2) a "design space" relating each treatment plan and household stratification to the structural model;

(3) an objective function, providing the measure by which the desirability of a design allocation is judged; and

(4) a total budget constraint and a vector specifying the cost per observation at each point.

Given the above information, the design problem is then to choose that distribution of households across design points which optimizes the objective function. Like the cost constraint, the objective function is capable of introducing factors which imply that nonorthogonality is a desirable feature. While its specific form may vary, in general we wish to minimize a weighted sum of variances associated with a vector of linear combinations of regression coefficients. The solution to the design problem specifies the number of households from each stratum to be allocated to alternative treatments; individual households are then randomly assigned according to the selection probabilities implicit in the solution.

If the Conlisk-Watts model begins with a correctly specified structural relationship, it can be a valuable tool in increasing the efficiency of an experimental design. It has been criticized, however, by those not wishing to let prior structural assumptions (which may be incorrect) condition the experimental design, and by the complexities it introduces in the use of experimental data for hypothesis testing. Some

of these issues will become apparent in the next section.

D. EMPIRICAL ESTIMATION AND HYPOTHESIS TESTING

The prototype model focused on the testing of a simple direct hypothesis concerning experimental effects. The experimental designs used for the income maintenance experiments were intended to accommodate more complicated hypothesis tests involving both variations in program characteristics and extrapolations to nonexperimental settings. The sample allocation was intended not only to permit the testing of these more complex hypotheses, but also to promote the precision of the intended tests.

While the samples for the NIT experiments were drawn according to the fundamental randomization principles necessary for applying conventional techniques of statistical inference, the design process was permitted to determine the choice and frequency of applied treatments and to choose probabilities of selection in alternative purposes and must be accounted for in designing methods of analysis.

The first point to be made is that the experimental design places limitations on hypotheses which can be tested. The New Jersey experiment was designed to vary controlled characteristics in a finite number of dimensions, and in such a way as to permit efficient testing of hypotheses related to a particular regression model. Alternative hypotheses may be tested, so long as the design space has sufficient dimensions to accommodate the tests. Such tests will be less efficient than if the experimental design had been established with those tests in mind. Thus the prototype design required hypothesis tests to be simple, but permitted more powerful tests of simple hypotheses than the designs used in the income maintenance experiments.

Secondly, the sample allocation process in the income maintenance experiments created a correlation between some household characteristics and form of program treatment. Thus simple bivariate relationships and comparisons of group means no longer have the direct interpretive value they would have had with orthogonal designs. For example, a simple test comparing mean earnings of families in a particular plan with those in the control group may be contaminated by the fact that family income influences the probability of assignment among treatments. That is, households in a particular plan and in the control group may be systematically different in terms of stratification characteristics, and simple group comparisons do not permit one to distinguish between the effects of plan and stratification variables on observed behavior.

This problem can be rectified by explicitly incorporating all stratification characteristics into the hypothesis test--either by controlling for all stratification characteristics in performing the test, or by making (and presumably defending) the assertion that the response being observed is independent of the stratification variables in question. Generally speaking, stratification of a sample by variables appearon the right-hand side of a regression model has no effect on the interpretability of coefficients or test statistics associated with that model.

The use of complicated experimental designs introduces definite risks that the prototype model avoided. It is necessary to specify a structural relationship between the behavioral response of interest and all variables used for stratification purposes in the design process. If this structural relationship is subject to specification error, the resulting experimental inferences may be incorrect. The prototype model, on the other hand, was more conducive to tests of experimental effect without knowledge of the underlying structure.

A third issue relates to projecting population estimates from results based on the experimental sample. The premise of conventional sampling theory--that a self-weighted random sample constitutes an unbiased representation of the population of interest--is not applicable to a situation where we induce behavioral responses in an experimental setting and requires an explicit theory for extrapolating to a nonexperimental situation.

Once we have confronted this situation, we may wish to translate measures of behavior for the experimental sample into unbiased estimates of what these measures would have been for a self-weighted sample of the population. The proper procedure involves a simple reweighting of the sample measures. There is a fundamental rule to be followed in this process, however, which is frequently violated: first estimate behavioral relationships on the raw sample, then reweight the distribution of point estimates where appropriate.

The reverse procedure of weighting observations prior to testing hypotheses, while equivalent for the direct calculation of variable means, results in incorrect estimation procedures in a regression framework. Consider an example in which labor supply is correctly specified as a linear function of the guarantee, the tax rate, and normal earnings, with a homoschedastic error term. Given these assumptions the appropriate estimation procedure, independently of how the distribution of observations by normal earnings corresponds to that of the population of interest, is unweighted least squares. To weight the observations would introduce heteroschedasticity in the error term and lead to an inefficient estimation procedure. If the error term in the raw regression is heteroschedastic, the weighting of observations and regressors (including the intercept) is an appropriate correction procedure, but these weights would bear no relationship to those involved in constructing population estimates.

Similar care must be taken in using experimental data sets for estimating behavioral relationships unrelated to the experiment. In particular, attempts to estimate behavioral

relationships involving stratification variables as dependent measures must utilize special estimation techniques.[10]

Finally, some general problems of statistical methodology related to hypothesis testing should be mentioned. In addition to the standard analytic problems associated with panel survey data--e.g., the need to deal with auto-correlated stochastic terms and with non-response bias and sample attrition--the fact that behavior has been experimentally manipulated creates special problems. Most of the experiments have been confronted with differential sample attrition rates by program status. Further difficulties are created when the structural models being tested require proxy variables such as "normal income," frequently essential in models of household economic behavior. On the one hand, it is hard to obtain a proxy free of induced experimental effects from the data for treatment households; alternatively, the construction of proxy variables from the same control group data used for making treatment-control comparisons can lead to small sample bias in constructing certain types of hypothesis tests.[11]

In summary, the designers of the income maintenance experiments deviated substantially from simple models in an effort to make the same design responsive to the structure of the hypotheses being tested. The cost imposed by this procedure was immense in terms of complexity imposed on the hypothesis tests and in terms of subtle analytic pitfalls created in the process. The net value of these design efforts is a subject of continuing dispute, with economists often taking a different position from observers in other disciplines.

E. EXTRAPOLATION TO NON-EXPERIMENTAL SETTINGS

The problem of extrapolation to non-experimental settings lies at the center of what makes the design of experimental samples different from the traditional practices of survey sampling. In survey samples we wish to obtain information about existing population characteristics without contaminating either behavior or household responses by the choice of survey methods. So long as such contamination can be avoided, well established random sampling procedures permit us to extrapolate sample characteristics to a total population of interest within known confidence intervals.

In a controlled experiment, on the other hand, an explicit attempt is made to apply stimuli to a sample of households in order to observe induced changes in behavior, and then to relate the results to the effects of applying similar stimuli to the total population on a non-experimental basis. The position taken by economists was that the prototype design model--with its comparison of independent "snapshots" of the population to measure experimental effects--was insufficient for handling the complex hypotheses to be tested. To them, experimentation meant exerting control over variations in program parameters and stratification characteristics--to permit estimation of structural relationships with random residuals. The emphasis of survey statisticians on random draws from populations and estimation of population means was of lesser importance.

The lack of correspondence between responses observed in experiments and program effects of direct policy interest comes from a number of sources.

First, the program to be ultimately considered for implementation is not known at the time experimentation begins, and is unlikely to correspond exactly to any of the treatments being experimented with. Thus, it may be necessary to extend experimental results to programs having similar but not identical characteristics.

Second, certain options considered for program implementation may not be viable subjects of experimentation. Since participation in social experiments is a voluntary process, the effects of policy options which leave some individuals worse off than under existing programs cannot be observed directly. Thus experimental results sometimes must be extrapolated beyond the scope of the tested programs.

Third, the results of an experiment depend both on the experimental program structure and on the environment faced by the control group. This background environment may differ from what is to prevail at the time of program implementation; thus it is important to standardize the environment of the control group where possible, and to understand its effects. Control of the background environment has proved to be one of the major problems in the social experiments. During the New Jersey experiment, for example, there were two major changes in the welfare system not only affecting the control group but also providing benefits more generous than those paid by some of the experimental treatments.

Fourth, certain features of an implemented program are virtually impossible to replicate or to observe in an experimental environment. The New Jersey experiment provided payments for only three years, while an implemented program would be of permanent duration. Some implemented programs--such as the transitional employment associated with Supported Work--would be similar in duration to their experimental counterparts, but their long-run effects may only be apparent after the results of the evaluation are required. Full scale implementation of a program may lead to different effects than those of a sample blown up to the full eligible population. For example, if the NIT had a major effect on the labor supply of low income households, it would have an impact on labor markets and wage rates unobservable in a small sample experiment.

Finally, there is the issue recognized from the very start of the experiments but not fully confronted: the possibility that individuals who are in an experimental setting may react differently than they would under normal circumstances.

Thus, the experiments do not provide direct answers to policy questions, but must be supplemented by nonexperimental analytic techniques. An integral part of the experimental process must be the provision of the information necessary for such analyses.

F. CONCLUSIONS

Prior to the advent of the social experiments, economists and other social scientists developed quantitative techniques for testing hypotheses with nonexperimental data. They developed methods of applied statistical inference which required prior acceptance of structural specifications. Econometrics became a well-developed form of applied statistics, and economists have long turned to their own profession for guidance in this area. The contribution of statisticians, on the other hand, could have been in the areas of sampling methodology and experimental design. While statisticians were consulted at various stages of the social experiments and made some valuable contributions, economists played a dominant role in design decisions.

During its early days, social experimentation was viewed as a technological revolution, and perhaps too much was expected of it. The social experiments are flawed in what they can do--not only because of errors in execution by economists and others--but also because creation of the appropriate experimental environment may be conceptually impossible. To be used correctly, therefore, experimentation must be viewed as an augmentation to existing methods of program evaluation rather than as a radical departure. Social experimentation exists today as a viable methodological tool because of economists and policy makers willing to listen to them; in the process it has acquired both the strengths and the weaknesses of their methodological perspective.

FOOTNOTES

1. The author wishes to thank David N. Kershaw and Cheri T. Marshall for their comments and contributions to the content of this paper. The views expressed here are the sole responsibility of the author.

2. For a review of the origins and design of the New Jersey Experiment, see Kershaw and Fair (5). See Rossi and Layall (8) for a major external critique of the experiment. Rossi is the leading critic of the experiment from a sociological perspective.

3. Ross (7).

4. Scranton, Pennsylvania was added to the original set of New Jersey cities after the New Jersey sample proved to be predominantly black and Puerto Rican.

5. See Hausman and Wise (3).

6. Portions of this and the following section draw freely from Metcalf (6).

7. Some economists have argued that experimentation with "extreme" treatments is useful in ways analagous to the use of extreme dosages in medical experiments.

8. See Conlisk (1) and Conlisk and Watts (2) for a discussion of the conditions under which orthogonality is desirable.

9. See Conlisk and Watts (2), Metcalf (6), and Rossi and Lyall (8) for detailed discussions of the Conlisk-Watts model.

10. In particular, see the discussion of truncated sampling frames in Section B above.

11. See Hollister and Metcalf (4) for a discussion of this issue.

REFERENCES

1. Conlisk, J., "When Collinearity is Desirable. Western Economic Journal 9 (1971): 393-407.

2. Conlisk, J. and Watts, H., "A Model for Optimizing Experimental Design for Estimating Response Surfaces." 1969 Proceedings of the Social Statistics Section, American Statistical Assn., pp 150-156.

3. Hausman, J., and Wise, D., "Social Experimentation, Truncated Distributions, and Efficient Estimation" in Followup Studies Using Data Generated by the New Jersey Negative Income Tax Experiment, Mathematica Policy Research, March 1976.

4. Hollister, R. and Metcalf, C.E., 1977. "The Labor Supply Response of the Family," in The New Jersey Income-Maintenance Experiment, vol. II: Labor Supply Responses. ed. by Harold W. Watts and Albert Rees, New York: Academic Press.

5. Kershaw, D. N. and Fair, J. 1976. The New Jersey Income-Maintenance Experiment, vol. I: Operation, Surveys, and Administration. New York: Academic Press.

6. Metcalf, C. E., "Sample Design and the Use of Experimental Data," in The New Jersey Income-Maintenance Experiment, vol. III. Expenditures, Health, and Social Behavior; and the Quality of the Evidence. Edited by Harold W. Watts and Albert Rees, New York: Academic Press (forthcoming).

7. Ross, H., "A Proposal for a Demonstration of New Techniques in Income Maintenance," Memorandum, December 1966, Data Center Archives, Institute for Research on Poverty, University of Wisconsin, Madison.

8. Rossi, P. H. and Lyall, K. C. 1976. Reforming Public Welfare: A Critique of the Negative Income Tax Experiment. New York: Russell Sage Foundation.

DISCUSSION

Bette S. Mahoney, Department of Health, Education, and Welfare*

Metcalf's paper is an interesting and useful one and there is much for a discussant to comment upon. Some of what he says applies to social experimentation in general and not just to the income maintenance experiments. There is much in it about which I will not comment and much with which I agree.

I disagree with the proposition that "the methodology of controlled experimentation has taken firm hold as the focal point of analyses of major changes in social programs." Experimentation is an important methodology for developing knowledge. It is both costly and limited in its results and total reliance upon it as the focal methodology, I think would be an expensive error. I see little evidence that such reliance has occurred. One has only to examine the recent Welfare Reform analyses for support of my contention. And since the proposition is not central to the discussion of the paper or the subject of the role of statisticians in experimentation, I will not discuss my concerns about social experimentation here.

The role of a statistician in social experimentation ought to be the design of the most efficient and effective methods to meet the purposes of the experiment. Metcalf describes a prototype model without regard to the purposes or the hypotheses to be tested. Because of this he presents a "strawman" prototype as the "contribution" of statisticians. One could expect statistiticians as well as economists to be more sophisticated in experimental design.

The paper's description of the difficulties in determining the "policy space" is useful. The additional questioning of the purposes of social experimentation is better suited for another discussion. The difficulties described in the Metcalf paper led to an innovative design, the Watts-Conlisk model, the merits of which are still being discussed. As Metcalf notes the design has its costs.

Metcalf calls our attention to the fact that the sample designes in the experiment are non-orthogonal and properly warns that the data therefrom must be analyzed with caution. I want to talk for a moment about this lack of orthoganality and the Watts-Conlisk sample allocation model which is responsible for it.

Cost, as Metcalf notes, is one important factor. Some observations cost more than others. The differences in cost raise two quite different problems. First, cost is not known a priori; it is, in fact, a major objective of the experiments to determine what the cost is. Thus, differences in cost per observation cannot be perfectly accounted for in the sample design unless one already has the knowledge that would make the experiment much less valuable, if not altogether unnecessary.

The second problem has to do with value, not with the experiments themselves (at least not in this context) but with the individual observations. One would not consider filling up the cells that are likely to be the cheapest unless one thought that an observation anywhere in the design space was equal in value to all the other observations.

Metcalf lists both cost and value among the four major components of the Watts-Conlisk model but does not dwell on them. He does comment about a third major component, the specification of the assumed structural relationship being tested. He observes that, if it is being properly specified, it can be a valuable tool in increasing the efficiency of an experimental design. He goes on to point out, however, that some social scientists have criticized the experiments for incorporating structural assumptions within the experimental design since those assumptions may turn out to be incorrect. And this leads me to my point: the assumptions about both the cost and the value of individual observations may also be incorrect and lead to errors in the sample allocation.

This is probably not the place to start a discussion about the value of knowledge. I will simply note that the Watts-Conlisk model assumed that policy makers were more likely to prefer some policy parameters than others. But to my knowledge, no one when using the model bothered to ask people running welfare programs which parameters were preferred before assigning weughts to the objective function. One result is that the experiments have generated virtually no information about the potential impact of plans with tax rates much higher than 70%. While many economists and other social scientists believe that policy makers should prefer plans with lower tax rates, there are others, including many administrators of current welfare programs, who disagree. They believe instead that plans with high tax rates are more efficient and effective.

Differences in value among different observations need not result only from the relative interests of policy makers. They may also be a function of how the data will be used. For example, suppose one important use of the data will be in estimating the cost of national programs. Other things constant, one would want greater relative estimating precision for plans with high tax rates than plans with low rates. This is so for exactly the same reason that high income taxpayers wish to be more accurate in estimating their annual income than low; to wit, the same relative error will be more costly. An objective function might specify that the dollar cost of plans with 50% tax rates and the dollar costs of plans with 70% tax rates should be estimable with equal absolute precision would assign more observations to the latter plan.

This might be the place to observe that data from the experiments have been used in estimating

the cost of the Administration's welfare reform plans. The data have also been used to buttress the contention that the labor withdrawal effects of the proposed plans will be within acceptable limits. But they played almost no role in the deliberations over which plan would be preferred and, as far as I can tell, which tax rate is to be preferred for its labor supply effects.

In fact, the choice of a 50% tax rate in the Administration's proposals are made despite the findings from the experiments of small labor force withdrawals and potentially higher program costs at these rates.

The administrative evidence which several years ago was felt to be of major importance to the implementation of a negative income tax has also been of limited use. One has only to compare sample sizes of 800 to 5,200 families in the experiments to 10 million households in the program to understand why.

There should be a role in the design of social experiments for the statistician. Hopefully it will be to do more than to design a "prototype model" like that described in the Metcalf paper. It seems likely that the constructive interaction of several disciplines might do more for the development of the methodology than the dominance by a single discipline even if it is my own.

As to the future of social experimentation, expensive as it is, I agree with Metcalf's view of it "as an augmentation" to other methods. As such, it can be very valuable but appropriate care should be taken to use it wisely.

*This discussion owes much to W. Michael Mahoney of the Social Security Administration for his invaluable assistance.

POOR MEASUREMENT OF THE RIGHT THING

Angus Campbell, The University of Michigan

The title of my paper comes from a statement by John Tukey, who in addressing the 1975 meeting of this society in Atlanta said, "It is often much worse to have a good measurement of the wrong thing than to have poor measurement of the right thing--especially when, as is so often the case, the wrong thing will in fact be used as an indicator of the right thing."

The groves of Academe and the humid banks of the Potomac are both crowded these days with people who talk about the quality of American life. They are all looking for the one right set of indicators which will tell us what the quality of life in this country is and whether it is improving or deteriorating. Until a few years ago this would not have been a great problem. After we learned to count the Gross National Product and the various income accounts that go with it, we only had to watch the GNP go up or down and we knew what was happening to the quality of life.

During the 30 years following World War II the GNP has generally been on an upward slope. Family income has increased by about two-thirds in constant dollars and the number of families living below the poverty line has dropped to about one in eight. These are important achievements; reducing the proportion of the population living in poverty is a national objective with which none of us would quarrel. But in increasing the number of families whose income is sufficient to purchase an adequate diet and the associated necessities of life, we have not increased the sense of confidence with which Americans walk the streets of their cities, we have not increased their feelings of security against unemployment, we have not strengthened the bonds which hold families together, and we certainly have not increased the citizenry's trust in their elected officials. Indeed it can be argued that as material welfare has increased in this country in the last quarter century, subjective well-being has declined.

It cannot be said of course that the Council of Economic Advisors and the other people responsible for the production of our economic indicators are unaware of the fact that a rising national income is not precisely the same as a rising sense of well-being. They inevitably come to talking about individual utilities if they carry their concept of social welfare to its logical conclusion. As my economist colleague, Thomas Juster, puts it, "The goods and services produced by the economic system, with rare exceptions, constitute instrumental rather than ultimate outputs of the system." The ultimate output is the subjective satisfactions and pleasures which flow from the supply of goods and services.

The problem seems to be not so much one of definition as of measurement. Economists are accustomed from the academic cradle to the use of data which have the quality of cardinality, data which permit them to insert a specific quantity of some unit (usually dollars) on the input side of an equation and predict or measure the output in the same unit on the other side. They are well aware that scales of satisfaction or happiness do not have this quality and they tend therefore to refer to subjective values as "intangible and unmeasureable." They undertake to locate indicators which can be easily counted which can serve as proxies or surrogates for what they consider to be unmeasureable; for example, the number of tickets sold to artistic performances of one sort or another might serve as an indicator of the public's level of aethestic pleasure, the number of vacation days taken might be used as an indicator of the total amount of the enjoyment of leisure, or the reported crime rate in a city or neighborhood might serve as an indicator of fear of crime among the residents of that area.

Although economists generally accept the proposition that the ultimate function of the economic process is to satisfy the needs of the population, their resistance to measuring these satisfactions directly can be said to be virtually total. A recent book review puts the issue in its bluntest terms:

> Seen from the point of view of economic theory subjective well-being is indistinguishable from the well-established concept of individual utility. After a century of discussions, we all came to know for sure that utility is non-measurable, noncomparable as between persons and nonsummable. There is no point in continuing to argue about that. Nonmeasurability implies not only that we do not know what scale to apply to the vertical axis in the utility diagram or to the third dimension of the indifference map, but also that the expressions 'very good, 'good,' 'satisfactory,' etc. used in respect of well-being positions have a meaning for separate individuals but not for interpersonal comparisons. There is no guarantee whatsoever that 'good' positions of various persons are in any meaningful sense equivalent. That fact alone is sufficient to undermine the whole concept of subjective well-being of a population. Concentrating efforts at the measurement of subjective feelings seems to be nothing but retreading of old paths which have proved to lead nowhere.

The problem for this reviewer and for everyone else who has thought about the problem is the difficulty of finding a common unit which will measure both objective products and subjective utilities. Dollars will not do. We have no way of converting so many dollars worth of product into so many units of utility. If we could find

such a common denominator the relationship between objective and subjective quality of life would become a matter of simple arithmetic.

In the absence of such a common unit the reaction of the reviewer I have been quoting is to dismiss the whole concept of subjective well-being. That does not mean that he has abandoned interest in the concept of quality of life; it means instead that he has decided that quality of life must be assessed in material terms because subjective measurement is impossible. In other words, in Tukey's language, he is prepared to use a good measurement of the wrong thing as an indicator of the right thing.

I do not think it is likely that we will find the magic numeraire that will solve the problem of converting products into utilities. There are also undoubted difficulties in the assumption that the utility one person assigns a product or an experience is directly comparable on a common scale to the utility another person assigns it, that an "interpersonal comparability of utility" is in fact possible. It may be argued that individual A and individual B may both say they get a great deal of satisfaction out of their work but in fact A's utility may be less than B's because A's expectations were lower than B's. The question then becomes which is more real to A and B, their sense of satisfaction with their work or their position on a scale of utilities that might be derived from their work. And even if we accepted the proposition that their sense of satisfaction is what is real to them we do not know precisely that a great deal of satisfaction feels the same to A as it does to B.

Generally speaking of course we are not concerned with clinical comparisons of individuals A and B but in comparisons of the social groups to which A and B belong. We are concerned with the quality of life of society as a whole and of its various segments. We would assume that the problem of individual variability in standards of judgment would be less serious when we are comparing large groups in which we would expect a certain amount of offsetting variation to occur. If we find that unemployed people are less satisfied with their lives than employed people of equivalent educational and occupational background we are inclined to believe that this represents something more than the vagaries of individual expectations. If we are able to follow these differences through time we are able to establish trends and to identify functional relationships which may exist between attitudinal and behavioral variables.

But there is no doubt that in comparison to the interval scales which are commonly used in counting economic products the ordered scales we use in assessing subjective utilities are weak measures. Our alternatives seem to be to use the established measures of economic products as our measure of quality of life and set aside the whole concept of subjective well-being (as our reviewer proposes) or to argue that subjective well-being is an indispensable attribute of quality of life and that the objective indicators measure it so poorly we are compelled to use the less precise subjective measures because they are at least attempting to measure the right thing.

In fact I think we have no alternative. As Robert Gordon recently observed in his presidential address to the American Economic Association, "Human welfare is a concept that will not go away no matter how uncomfortable it makes the economic theorist." In a society as politically free as ours it is impossible to imagine that the public's sense of well-being or discontent can be ignored. Values which cannot be accounted for in a traditional economic balance sheet are important to people and influence their behavior. They include the enjoyment of social relationships, the satisfaction of challenging work, the respect of friends and associates, a sense of security from attack in their homes and on the street, peace and quiet in their neighborhoods, pleasure in the appreciation of natural beauty and many others. Few people have abandoned interest in the economic realities of life but their lives are not as preempted by economic considerations as the conventional image of economic man might have led us to expect.

However many reservations we may have about our ability to measure these subjective utilities there is no doubt that policy-makers in a democratic society have to be concerned about them. They may very well draw up a balance sheet which gives them a detailed statement of the economic costs and benefits to be expected from a specific policy. But, whether explicitly or not, they also have in mind a second set of accounts where the utilities and disutilities to be expected are entered. The fact that these latter entries may be based on imperfect evidence does not make them insignificant. They may be imprecise measures but they are indicators of something the policy-makers recognize as important.

We are currently witnessing an example of this double bookkeeping in the controversy over the admission of the Concorde aircraft to Kennedy Airport. One set of accounts will show the financial benefits to the City of New York, the time saved by busy passengers, and other objective gains to the New York community. A second set of accounts will record the annoyance of New York residents with the noise associated with Concorde overflights. The people who make the decision at Kennedy will consider both of these sets of accounts and it may well be that their evaluation of these conflicting indicators will be as much influenced by the subjective factors as by the objective. They will not need to be able to convert annoyance into dollar amounts nor will they be much concerned whether one annoyed person has exactly as great a disutility as another. They do know that an annoyed electorate is capable of expressing its resentment and that public officials who disregard the public's sense of well-being and ill-being do so at their own peril.

A society as committed to the values of human rights and civil liberties as ours is cannot hope to represent the quality of its national life adequately by counting the usual economic and sociological indicators. The Eastern European countries lean heavily

on their statistics on employment, medical service and educational enrollment as indicators the quality of their lives; they do not talk much about nonmaterialvalues. But we must take account not only of the objective circumstances in which our people live but of the desirable and undesirable impact these circumstances have on their life experience. Monitoring our rates of crime, divorce, abortion, unemployment, pollution and disease undoubtedly tells us something about this experience as do statistics on leisure time, vacation travel, participation in artistic events and other such positive episodes of life. But it must be clear that these indicators however countable they may be, are only inferential and very partial.

Our economic indicators tell us that for the last 30 years we have had a rising standard of living with an associated increase in educational achievement and professional and technical employment. The very fact of these trends makes these indicators less capable of giving us an adequate description of the quality of American life. A growing proportion of our people are being liberated from a preoccupation with income, their horizons are being extended, the awareness of alternatives raised, and their concern with noneconomic values increased. There is no doubt that we should extend and refine the accounts we keep on standard of living and the objective circumstances of life. They tell us a great deal and they are indispensable. But we will need a different set of accounts to inform us about the subjective experience of life. They will not be as precise or as elegant but they will be measuring the right thing.

IS SATISFACTION A GOOD MEASURE OF THE PERCEIVED QUALITY OF LIFE?

Tom Atkinson, York University

In May of 1976, the Canada Council announced the award of a major long-term grant for the study of subjective or perceptual social indicators to a group of researchers associated with the Institute for Behavioural Research at York University in Toronto. The grant, which covers five years of research and provides about $1 million, was the first large award for empirical research on subjective social indicators in Canada. This paper will provide an outline of the research design and its rationale and go on to discuss some of the measures being used to assess what has become the central subjective social indicators - the perceived quality of life.

Overview

Between 1970 and 1975 a substantial amount of research was initiated in the United States and Britain on subjective social indicators -- that is, measures of personal perceptions, preferences,attitudes, values, etc. At the Center of this work was the research of two groups at the University of Michigan - Angus Campbell, Philip Converse and Willard Rodgers in one and Frank Andrews and Stephen Withey in the other. Both groups worked from a common conceptual model but differed in the types of measures which they preferred and in the purposes of their projects. In the United Kingdom, Mark Abrams and John Hall undertook a series of studies which shared a conceptual model and methodology with the Michigan work, particularly the Campbell, et. al. formulation. The focus of all of these efforts was perceptions of the quality of life -- the subjective indicator most directly analogous to the objective quality of life concerns underlying the work in the OECD internationally and in government departments such as HEW in the United States.

Elsewhere in the U.S., research on subjective indicators has been undertaken by the Survey Research Centre at Berkeley which was primarily concerned with prejudice and alienation and at the National Opinion Research Center with their General framework of research on subjective social indicators.

In summary, during this period the extensive funding for research on subjective social indicators indicated that it was an idea whose time had come. Further, the extent of research in the area led to the conclusion that many of the ticklish measurement problems had been, or were about to be, resolved. The rather luke-warm reception given the products of this research by funding agencies and by other social scientists, at least in the United States, was not yet apparent nor were the flaws in the research that led to such a response.

The Canadian subjective indicators project on which we are now embarked drew its initial inspiration from the work at Michigan, particularly from earlier papers by Campbell and Converse, and from the part of their research which dealt with the role of what they called "standards of comparison", that is, levels of expectation, aspirations and other comparison points used in evaluating any situation or object. It took as its starting point the conclusion that the American and British efforts had successfully evolved measures of the perceived quality of life and these measures could now be used to develop social indicators measured over time and on a national basis.

Our research has two major objectives. The first is to develop several subjective social indicator measures which can be used to describe the national population and subgroups within it. These measures, although derived from cross-sectional surveys, were to be collected over time to develop indicators of change. The second objective was to examine the causal agents responsible for variation in these indicators across the nation and over time.

It's this latter objective which holds the greatest promise for the development of social indicators in general because it leads to an examination of the ties between objective and subjective social indicators. One of the potential uses of subjective indicators research is that it can inform the development of objective systems by identifying those objective indicators which have a significant impact of perceptions of the quality of life. Without such a test of relevance, the creators of social statistics have no criteria for deciding which of the multitude of objective indicators should be included in a system of social indicators. By establishing covariance relationships between objective indicators and their subjective counterparts, efforts can be focused on the generation of highly accurate, spatially-detailed statistical information systems which can be used to produce summary measures of demonstrated importance to the population's perceptions of the quality of life.

While some attention has been paid to the relationships between subjective indicators and the objective conditions,most of those efforts have been directed at data which can be collected via self-report,such as income. As a result, little analysis of the effects of the economic, social, political and physical attributes of the local environment has been conducted. The lack of enthusiasm for most recent research on subjective indicators may stem, in part, from the absence of aggregate objective measures which form the core of most objective indicator research. This shortcoming is not inherent to investigations of subjective indicators but to deal effectively with it requires a research design which incorporates data on geographical areas as well as on the subjective responses of individuals living within those areas. The design should also be influenced by the desire to investigate the objective-subjective links over time since an analysis which relates changes in one to the other is more powerful than one limited to a single point in time.

Another area which has not received adequate attention in research on subjective indicators is the investigation of the perceptions and attitudes of elites in the government and private sectors. Elites are important in the context of social indicator research for two reasons: first, they and

their decisions both influence and are influenced by public perceptions and attitudes regarding the quality of life and other subjective indicators; and, second, elites are often the leading edge of social change in that,through a variety of mechanisms, their preferences and prejudices often are strong influences on the direction of social change. This latter statement may be overstated because we know very little about the impact of elite dispositions on the direction of social change particularly in a highly-decentralized social democracy such as exists in Canada, where considerable conflict may exist among elites with differing goals and values. In fact, changes in elite attitudes may follow rather than lead changes in the general public rather than the reverse but it is the uncertainity about the size and direction of these effects which recommends them as research topics.

It is however, the perceptions of elites in different levels and their relationships to public perceptions which are of central interest to social indicators research. Elite perceptions of the quality of life in different areas and their perceptions of the public's level of satisfaction in those areas influence the types of policy which will be endorsed and the content of messages which may be transmitted, via the media, to the public. To the degree that public and elite perceptions and attitudes are consistent and that elites are consistent across sectors, actual or potential social and political conflict is lessened and the direction of social change becomes more apparent.

It is not clear that we will be able to resolve most questions about the complex connections between elite and public preferences or among sectoral elites in the course of a five-year study. It is clear, however, that if social indicators research is interested in doing more than describing social change after the fact, it must incorporate elite research with the type of studies of the public now being undertaken.

Our general research schema is represented in Figure 1.

I have briefly discussed the rationale for the concern with environmental characteristics and elite behaviour as determinants of subjective indicators. The "Life Events" component constitutes the third major cluster of causal variables in that significant personal events, such as marriage or divorce, job advancement or loss, changes in family size and so on, have a large impact on perceived life quality and other subjective indicators. These events result, in some cases, from changes in environmental characteristics and in other cases are independent of them -- for example, changes in life state which result from aging. Any study which attempts to identify the major agents responsible for changes in subjective indicators should examine the role of life events both as mediators of a changing environmental conditions and as independent causes.

Research Design

The discussion, to this point, has been concerned with general research objectives and an overview of the major clusters of variables. I would like to now turn to the specific research design being implemented to generate the data required to examine the critical relationships. Four major data collection activities are underway with a fifth to be undertaken at a later date. As shown in Figure 2, they are:

1. Cross-sectional surveys. National surveys of Canadian population will be undertaken in 1977, 1979 and 1981 to develop time-series measures of the central subjective indicators. Samples of two thousand respondents will be regionally stratified to produce fairly accurate regional estimates as well as very accurate ones for the national population. Although the main purpose of these surveys is to develop good descriptive data, the sample selection procedures are designed to hold geographical areas, in this case Census Tracts, constant across surveys. Given the survey design only 160 of the approximately 4000 in Census Tracts in Canada will be sampled but the same 160 will be included in each wave of survey work.

The constancy of these geographical units allow us to develop measures of the environmental characteristics in those areas and relate changes in them to measures of subjective indicators. The design of the cross-sectional survey reflects both our desire to develop good descriptive subjective indicators and to examine the causal linkage between objective and subjective measures. The key to untangling those connections, at least in this study, is the five-year duration of the investigation which permits the analysis of covariation over time.

2. Panel Surveys. Panel surveys in two cities will be done in conjunction with the national surveys in 1977, 1979 and 1981. One thousand respondents, evenly divided between the Toronto and Montreal metropolitan areas, will be interviewed. Unlike the cross-sectional survey in which the geographical areas remain constant while the respondents change,the panel holds constant the respondent while not constraining geographical location. Given the mobility rates in these two cities, it is anticipated that 50% of the panel will move within the five-year duration of the study. This component of the research provides us with an opportunity to investigate the effects of changes in environmental characteristics and life events on perceptions of the quality of life and other indicators.

Toronto and Montreal were selected as panel locations because a) over 20% of the Canadian population lives in the two cities, b) they are easily accessible to the project research group which can independently develop objective measures to supplement the data available through government agencies, and c) both provide highly varied urban environments -- some of which are very stable while others are subject to rapid change.

Since the primary purpose of the panel survey is to investigate the dynamics of subjective indicators rather than produce representative descriptive measures of the urban populations, a procedure for selecting panel members which insures the inclusion of those likely to experience change in thier lives will be utilized. During the first wave of survey work, the cross-

sectional and panel respondents in Toronto and Montreal were combined producing approximately 700 interviews in each city. Respondents will then be selected for reinterview in the panel so as to maximize the occurance of those who have a high probability of change in housing, job and family composition since the initial interview.

3. Elite surveys. Elite surveys will be conducted each of the three years in which the public is surveyed. The elite sample is selected positionally -- that is, positions within sampled organizations are selected and the incumbent interviewed. In most cases the senior administrative officer is selected from organizations in the private sector while senior elected officials and civil servants are included from government agencies. The sample is composed of 550 respondents drawn from the following areas: large corporations, small business, labour unions, government (elected and civil service positions from federal, provincial and local levels), the legal profession, media, agricultural organizations and the academic community. The largest segments of the sample will come from the corporate and governmental sectors.

The elite sample is designed as a panel that is defined by position rather than person. Given the normal rate of turnover in these senior positions, it should be possible to distinguish the effects of role or position on elite perceptions. As a result of the over-time aspect of the study, the sensitivity of elites to changes in environment and in public attitudes can also be examined.

4. Ecological Data Base. This data base is composed of statistical information on the econonic, social political and physical attributes of the geographical areas in which the respondents live. It includes indicators from each of the areas one usually finds in volumes on objective social indicators -- health, employment, safety, housing and so on. These measures are, in most cases, available through governmental statistical services but additional indicators may be developed by the research group in the Toronto and Montreal areas.

The organizing unit for this data base is the Census Tract because it most clearly parrallels the idea of neighbourhood. It has been argued by Rossi and others that many contextual variables manifest themselves most clearly at the neighbourhood level. We would expect, for example, neighbourhood crime rates and population densities to be more closely tied to the perceived quality of life than measures of those variables compiled for the city or metropolitan area. Other types of indicators such as cultural facilities, job vacancies and cost of living measures may be more appropriately developed for larger aggregates. The ecological measures will generally be included for the smallest aggregation for which they are available. Special tabulations may be required to produce indicators at the appropriate levels in large urban areas.

5. Media Content Analysis. Although currently scheduled as a future project, we intend to develop a content analysis of daily newspapers in the ten major Canadian cities and news programs on the two national television networks. This information should give us some understanding of the manner in which the media filters information between the public and the elite.

The research design is quite complex and ambitious but each element is required if we are to pursue our dual objectives of developing good descriptive subjective indicators at a national and regional level and exploring the factors responsible for variation in those indicators.

Measuring the Perceived Quality of Life

The central subjective social indicator in this project is perceived quality of life. Drawing from the research at Michigan, we focused on measuring the perceived quality of life in general and in specific areas by asking the respondents to evaluate their own lives using identical measures across all areas. Andrews and Withey have shown that evaluations of a small number of areas or domains can capture most of the variance in perceived quality of life. Our misgivings about the conceptual independence of the central domain of the Andrews work -- evaluations of self -- led us to drop that particular area but we have used most of those identified by Andrews and Withey and used by Campbell and his colleagues.

The major controversy in this research, however, does not involve what areas or objects are to be evaluated but what measures are best suited to the task. Four types of measures have been suggested to tap perceived quality of life: a)cognitive measures such as satisfaction used by Campbell, Converse and Rodgers, and by Abrams in England, b) affective measures such as happiness used by Bradburn and in the Gallup Poll, c) measures which combine the two such as the Andrews-Withey Delighted-Terrible scale and d) self-anchoring measures such as Cantril's Ladder Scale and George Gallup's modification of it -- the Mountain Scale.

Of the four, satisfaction measure and the Andrews-Withey measure have received the most attention, and I will consider them here -- saving a discussion of the self-anchoring scale for later in this paper. The difficulty in deciding among measures can be clearly understood when one realizes that two very competent groups of researchers working out of the same research institute at the University of Michigan did not arrive at the same measure of perceived quality.

The Campbell research, which was conducted earlier than Andrews', utilized a seven-point satisfaction-disatisfaction continuum to measure perceived quality. Their choice was consistent with that psychological adage that a seven-point scale is all that most individuals could deal with effectively. Whatever the reasoning, the use of this scale proved the Achilles heel of their research. The difficulty with the measure resulted from a very serious skew toward the positive end of the scale. In all fifteen specific domains which were assessed, the modal response to this scale was the highest one -- "Completely Satisfied". In twelve of the fifteen over one-third of the sample indicated complete satisfaction with their life in that area. The general

satisfaction scale was not quite as positive with 22% indicating complete satisfaction and an additional 40% in the adjacent category.

These highly positive distributions had two negative consequences: first, there was so little variance in the measures that the investigators were forced to present almost all of their data as standardized scores thus eliminating any comparison of absolute scores over time, and second, these data flew in the face of the assumptions held by many academics, policy-makers and social commentators who maintained that the quality of life in America had declined in recent years. Since the satisfaction scores of disadvantaged groups such as Blacks and the poor were only slightly lower than others, many researchers concluded either that satisfaction was a poor social indicator or that the measures used were flawed. As a result, the Campbell, et. al. study has had little impact on the direction of social indicators development and has not encouraged funding for additional research.

The 7-point scale developed by Andrews and Withey is not, strictly speaking, a satisfaction measure and represents an attempt to "improve" the shape of the response distributions. They have mustered an impressive body of evidence to demonstrate that the Delighted-Terrible scale does, in fact, reduce the proportion of respondents in the top category while maintaining the size of the correlations among the various domain measures and with demographic variables such as income. In addition, they have shown that the measure is relatively free of method bias.

There are, however, three difficulties with the Andrews-Withey scale from our point of view. The first and most serious was that the variance of their measures was, in may cases, lower than the satisfaction measures used by Campbell. The second was that we wanted to experiment with expanding the scale and it would have been difficult with a scale composed of emotive words. Finally, national studies in Canada are conducted in English and French and difficulties of translation could easily destroy the comparability of the measures. Because of these difficulties we decided to focus our efforts on rectifying the satisfaction measure rather than use the Delighted-Terrible scale. It was clear that the satisfaction measure as used by Campbell was in need of modification contrary to our initial premise that measurement problems had been resolved.

The most direct suggestion for modification came from the British Quality of Life research which began by using a seven-point satisfaction-dissatisfaction measure and dropped it in favour of an eleven-point scale in 1973. They have not presented a rationale for the change but the response distributions indicated greater variance and less top-end loading with the longer scale. The two British surveys incorporating different versions of the measure were separated by two years and not directly comparable but their results encouraged our speculation that scale length was an important variable.

During the past year we have conducted three pretests which included different satisfaction measures all with identical question wordings. Each pretest was conducted in Toronto and Montreal and was evenly divided between English and French respondents. The sampling procedures for Pretests A and B were comparable but somewhat different from Pretest C so that B and C should not be directly compared. Table 1 shows the distributional attributes of the scales and indicates the effects of lengthening the scales. The criteria for evaluating these figures are not well established but increasing scale length seems to clearly improve several scale attributes. Variance increases with length while the proportion of respondents in the highest category decreases. Skew and Kurtosis decrease or remain in the same range. A comparison of the seven-point and eleven-point scales used in Pretest B shows that the top two values in the eleven-point version contain the same proportion of respondents as the "Completely Satisfied" response of the seven-point scale. These data suggest that, at the top end of the scale, the seven-point scale unnecessarily compresses the distribution and overstates the segment of the sample which is completely satisfied.

There is the possibility that much of the variance introduced by the longer scale length is random variation. One method for evaluating that possibility is to examine the correlation of each scale with a criterion variable. If the variation is random, the correlations using the longer scales will be significantly lower than the short scales. Unfortunately no criterion variable is possible when dealing with subjective variables of this sort but we can compare the correlations between the financial satisfaction measures and income. Those correlations are: .16 for Pretest A which used the five-point scale, .40 for the seven-point scale in Pretest B and .42 for the eleven-point scale in that pretest, and .23 for Pretest C with the eleven-point version.

There is no evidence in these figures to support the contention that the increase in variance obtained with the eleven-point scale is random variance. Comparisons of seven and eleven-point scales in other domains are consistent with this interpretation as well. On the basis of these analyses, we have concluded that an eleven-point satisfaction measure is preferable to the seven-point scale used by Campbell and his colleagues and to the seven-point Delighted-Terrible Scale developed by Andrews and Withey. It may not, however, be superior to an eleven-point version of this latter scale but the problems incurred in the expansion of the scale and its translation into other languages seem insurmountable.

Before tackling the last issue of this paper, some brief speculation about the reasons for the differences between these two scales is appropriate. Respondents seem to determine their answers to scales with positive and negative poles through a two-step process. First, they decide if they are positive, negative or neutral about the issue, and then they determine the degree of positiveness or negativness. Thus a seven-point scale is, in effect, a three-point scale in this latter step while an eleven-point scale is a five-point scale.

Respondents also seem to divide the response continuum on the positive or negative side into roughly equal proportions according to the number of scale values. As a result of this division, the value identified as "Completely Satisifed" covers a larger range of responses as the scale length decreases. To suppose that a value labeled in such a way has an absolute meaning outside of the choice context in which it is presented ignores the psychological research which shows that an individual's choices vary with the options presented since the information conveyed in the alternatives helps define the meaning of each choice.

Is satisfaction a measure of perceived quality of life?

Almost all of the research on the perceived quality of life in the United States and England has focused on satisfaction or satisfaction-like measures such as the Delighted-Terrible scale. I want to contend that, in one sense, these are not measures of the perceived quality of life -- rather they are responses to the perceived quality of life. Satisfaction measures result, in large part, from the comparison of aspirations and expectations with one's current situation. Thus it is possible, if not probable, that individuals could assess their quality of life as high yet be dissatisfied and as average or low and be satisfied.

It is these potential discrepancies between perceived quality and satisfaction that lead some policy analysts to write-off subjective indicators like satisfaction because they feel that the poor or other disadvantaged groups are too often satisfied with bad lot while the middle and upper-classes are discontent with a good one. I am not arguing here that satisfaction measures have no place in subjective social indicator research but that other measures, which may be closer to the perceived quality of life concept, have been neglected.

Figure 3 shows the Campbell, et. al. model of satisfaction and modification of it that follow from my argument. The initial model holds that the perceived attributed is compared to some standard such as level of aspiration and an evaluation arrived at which is level of satisfaction. The extension of the model inserts a prior assessment of quality which results from a comparison of the perceived attribute with some standard of excellence perhaps defined by what others have. In concrete terms, the difference can be illustrated as follows: the first model suggests that a man and his family living in a three-bedroom house with one bath might say to himself that it had always been his ambition to live in a house with four bedrooms and two baths and determine that he was dissatisfied with his housing. The expanded model indicates that he would arrive first at an assessment of whether his housing was of good quality or not and then, compare it with his aspirations and expectations to determine whether it was good enough to be satisfactory or bad enough to be unsatisfactory.

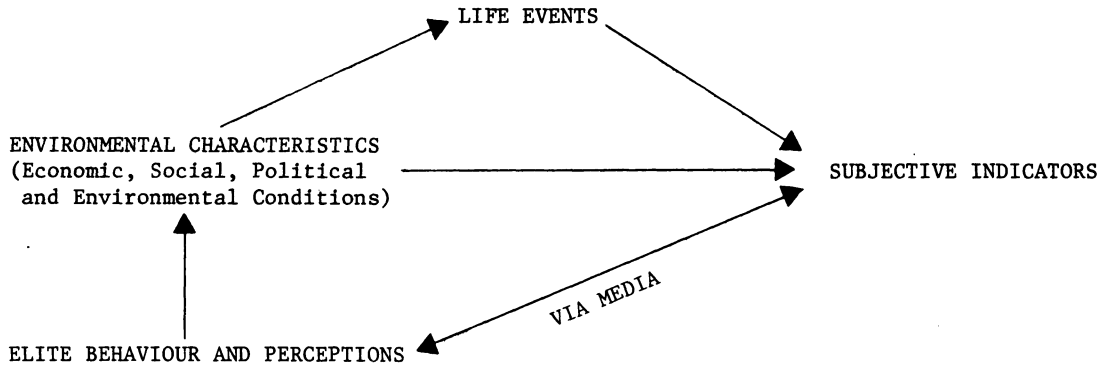
The major problem resulting from this argument is that even if we believe that satisfaction and perceptions of quality are conceptually distinct, how do we measure each independently. Research on consistency theory in psychology indicates that if the two evaluations are inconsistent there will a pressure to revise one or both to produce a better match. We have attempted to develop a measure of perceived quality by using the self-anchoring ladder scale shown in Appendix A and, more specifically, by comparing the respondent's assessment of his own position and that of the average person living in Canada. The difference or gap score is not a "pure" measure of perceived quality but I would argue that it is a measure of perceived quality relative to a specific reference group and is closer to the quality of life concept than satisfaction. Table 2 shows the distribution of the satisfaction and ladder scores for financial situation in two of the pretests.

Of greater interest are the correlations among these measures and between them and income shown in Table 3. Both matrices show the expected high correlations between the satisfaction measure and the ladder rating of financial situation. These correlations are enlarged somewhat because of correlated methods effects. Correlations between satisfaction scales and difference measures (.46 and .56) are a better indication of the relationship between perceived quality and satisfaction because they are not subject to common method variance. Family income shows a higher correlation with the difference measure than satisfaction in Pretest C as we would have predicted but the reverse was true in Pretest B.

If difference scores derived from the ladder scales measure a construct which is at least partially independent of satisfaction, then we would expect the correlation between the difference measure and income to remain when the effects of satisfaction were held constant. This is, in fact, the case as the correlation in Pretest B was reduced from .33 to .17 ($p < .05$) and in Pretest C from .34 to .26 ($p < .01$)

The existence of these independent relationships has encouraged attempts to pursue at least two types of subjective indicators -- perceived quality and satisfaction. The difference measures seem to approximate the former, although we do not yet know enough about how they work, and the eleven-point satisfaction scale looks like a good measure of that variable.

We hope that this research will inform and encourage the efforts of other as the research at Michigan and in England have benefited and encouraged us. There is no other area of social research that offers greater need for our possibility of international cooperation than the social indicators area. Let us learn from each other's success and failures.



INFLUENCES ON SUBJECTIVE INDICATORS

FIGURE 1

	CROSS-SECTIONAL SURVEY	PANEL SURVEY	ELITE SURVEY	ECOLOGICAL DATA BASE	MEDIA CONTENT ANALYSIS
Primary Purpose	Develop a range of subjective indicators at national and regional levels and measure them over time.	Investigate the causes of variation in subjective indicators, particularly the effects of objective conditions.	Measure subject-indicators for elites from different sectors and assess their perceptions of the public's levels of satisfaction and quality of life.	Organize data on the objective characteristics of the local environments in which the survey sample resides.	Code the contents of major daily newspapers as they relate to quality of life domains.
Scope	National, regionally stratified.	Toronto and Montreal Census Metropolitan Areas.	National with provincial and local elites.	National.	National, major cities.
Data Collection	1977, 1979, 1981	1977, 1979, 1981	1977, 1979, 1981	1975-1981	1977-1981
Sample Size	2000	1000	550	160 Census Tracts nationally, with an additional 83 in Toronto and Montreal	About 20 daily newspapers.

COMPONENTS OF PROJECT

FIGURE 2

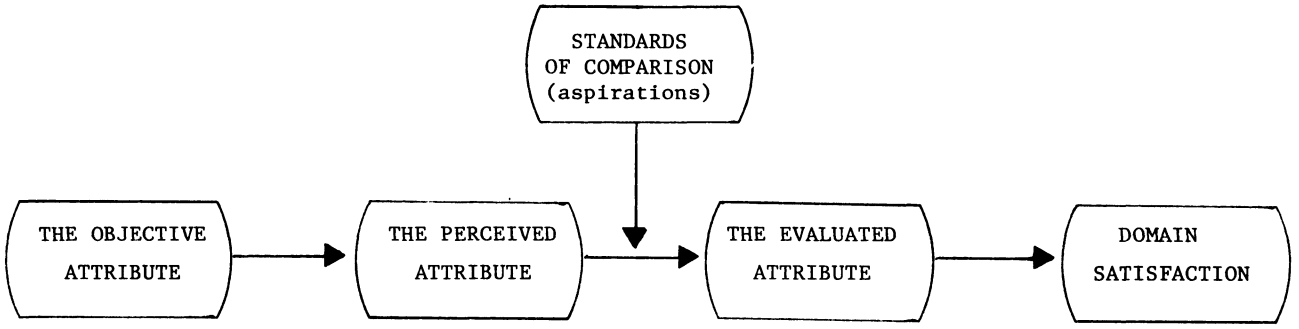
		General Satisfaction				Financial Satisfaction			
		Pretest A	Pretest B	Pretest B	Pretest C	Pretest A	Pretest B	Pretest B	Pretest C
Highest Score	11	*	*	6%	6%	*	*	5%	4%
	10	*	*	11	16	*	*	5	5
	9	*	*	15	23	*	*	8	17
	8	*	*	18	17	*	*	11	21
	7	*	17%	9	14	*	9%	11	18
	6	*	28	17	12	*	16	18	12
	5	19%	24	7	7	12%	19	8	7
	4	69	19	7	1	36	20	11	9
	3	9	7	3	3	29	18	10	4
	2	2	4	3	1	19	11	7	1
Lowest Score	1	1	1	3	0	4	8	8	1
Mean		4.03	5.12	7.13	7.86	3.33	4.17	5.69	7.07
Standard Deviation		.67	1.42	2.45	2.04	1.05	1.72	2.75	2.12
% Highest Category		19	17	6	6	12	9	5	4
% Two Highest Categories		88	45	17	22	48	25	10	9
% Below Midpoint		3	12	23	12	23	37	44	22
Skew		1.23	.60	.52	.69	.26	.13	.04	.46
Kurtosis		4.13	.12	-.30	.09	-.64	-.87	-.85	-.22

* = Scale value not included

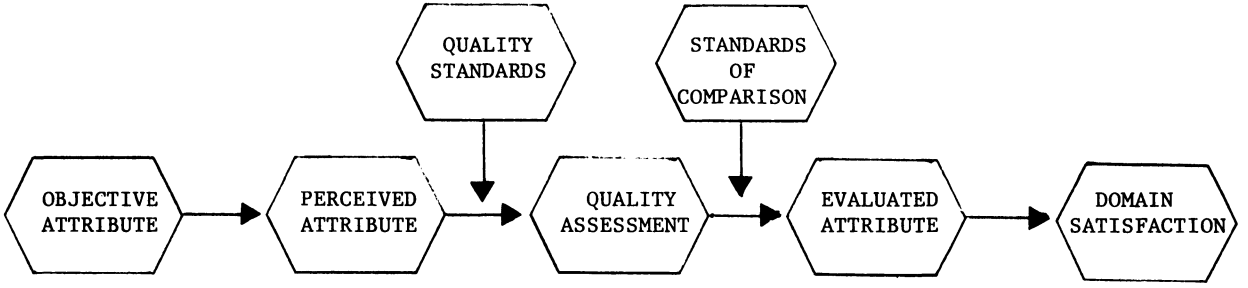
Attributes of Satisfaction Measures

Table 1

Campbell, Converse and Rodgers Model



Revised Campbell Model



TWO SATISFACTION MODELS

FIGURE 3

		Pretest B			Pretest C		
		Satisfaction 11-Point	Ladder/ Self	Ladder/ Average	Satisfaction	Ladder/ Self	Ladder/ Average
Highest Score	11	5%	1%	3%	4%	1%	1%
	10	5	1	1	5	3	2
	9	8	7	6	17	21	12
	8	11	9	14	21	17	26
	7	11	18	25	18	20	26
	6	18	27	29	12	19	22
	5	8	16	18	7	8	7
	4	11	11	3	9	6	1
	3	10	6	1	4	1	0
	2	7	2	1	1	1	0
Lowest Score	1	8	3	1	1	5	3
Mean		5.69	5.90	6.55	7.07	6.78	6.97
Standard Deviation		2.75	1.93	1.57	2.12	2.15	1.71
% Highest Category		5	1	3	4	1	1
% Two Highest Categories		10	2	4	9	4	3
% Below Midpoint		44	38	24	22	21	11
Skew		.04	.13	.19	.46	.93	1.25
Kurtosis		-.85	.26	1.46	-.22	.72	3.15

Attributes of Financial Situation Measures

Table 2

Pretest B

		1	2	3	4	5
1.	Satisfaction	-				
2.	Ladder/Self	.65	-			
3.	Ladder/Average	.15	.18	-		
4.	Ladder/Self-Average	.46	.73	-.54	-	
5.	Family Income	.42	.32	-.07	.33	-

Pretest C

		1	2	3	4	5
1.	Satisfaction	-				
2.	Ladder/Self	.63	-			
3.	Ladder/Average	.01	.45	-		
4.	Ladder/Self-Average	.56	.67	-.37	-	
5.	Family Income	.23	.46	. 13	.34	-

N = 150 $p > .05 = .16$, $p > .01 = .21$

Correlations Among Financial Situation Measures

Table 3

THE STRUCTURE OF SUBJECTIVE WELL-BEING IN NINE WESTERN SOCIETIES[1]

Frank M. Andrews and Ronald F. Inglehart
Institute for Social Research, The University of Michigan

1. INTRODUCTION

Interest in social indicators of life quality, including citizens' perceptions of their own well-being, has inspired a number of sample surveys in recent years.[2] Such surveys, particularly when done on a comparative and repetitive cross-national basis, have enormous potential for providing information about changing levels of social and economic development and about the processes and conditions that lead to or are associated with the "good life." However, the feasibility and usefulness of comparative research in this area--as in any area--are contingent upon the identification of an underlying phenomenon that is in fact comparable from one society to another. While a person's sense of happiness, satisfaction, etc. is of acknowledged importance, the cross-cultural comparability of the phenomenon of perceived well-being is largely unexplored. This paper reports an initial, and necessarily incomplete, examination of the comparability of psychological structures of subjective well-being in nine western societies.

This Introduction develops the conceptual framework for the analysis that follows, and describes some of the interests that motivate this presentation. Section 2, Data, describes the sample surveys from which reasonably comparable data from nine nations have been extracted and details the items and response scales used to measure perceived well-being. The section on analysis methods discusses the statistical techniques by which we identified the structures of perceived well-being and assessed their similarity across countries. There follow the main substantive results--first for USA and then for eight European nations. The final section of the paper provides some general conclusions, some cautions about interpretation, and some suggestions for further investigation of the issues.

Research on perceived well-being commonly distinguishes between evaluations of life-as-a-whole (sometimes referred to as general or global evaluations) and evaluations of specific life concerns, such as housing, job, relations with other people, safety, or fairness. When we refer to the "structure of subjective well-being" we refer to the way specific life concerns, and evaluations of them, fit together in people's thinking. For example, we ourselves have shown that among American adults evaluations of one's marriage are--quite reasonably--strongly related to evaluations of one's spouse, that evaluations of national political leaders are strongly related to evaluations of government economic programs, but that evaluations of the first pair are virtually independent of evaluations of the second pair. These statistical results suggest that Marriage and National Government are distinct life concerns for most Americans. When one combines these results with numerous others, some of which will be described later in this paper, one can identify a psychological structure, or "cognitive map," from which one can infer the relative positions of life concerns as they are perceived by a particular group of people.

Such structures are interesting for a number of reasons. In showing how well-being perceptions are organized in people's thinking, they indicate some fundamental aspects of what evaluations of life quality mean to these people. Such structures help to identify the distinct well-being concerns that particular groups have, and show the extent that evaluations of these different concerns overlap or intersect with one another. This suggests one of the important practical uses of such structures: They provide guides to the adequacy of coverage and statistical efficiency of indicators of perceived well-being. To the extent that people in different societies organize their thinking about well-being in basically similar ways, it is feasible and potentially productive to undertake cross-cultural research with standardized instruments and to make well-grounded comparative statements based on the results. However, if the basic phenomenon that is being investigated--well-being perceptions--shows markedly different structures in different societies, measurements and interpretations must be society-specific and any comparative statements must be advanced with extreme caution.[3]

The main substantive purpose of this paper is (a) to explore the structural similarity of well-being perceptions in nine western societies. In so doing, we shall have the opportunity to pursue two other matters of more didactic interest. (b) Our analysis is based on a set of national sample surveys that offer rich opportunities for secondary analysis, and our use of these data may increase analysts' awareness of their existence and accessibility. (c) This analysis involves use of some relatively new statistical methods for assessing similarities among configurations (i.e., structures) and illustrates the need for some further statistical developments; perhaps it will encourage statisticians to pursue these developments.

Before proceeding further, the reader should be cautioned that the analysis reported here is of a rather exploratory nature. The issue of cross-cultural similarities in structures of perceived well-being is a fundamental one for those interested in comparative research or in social policies, but the data requirements for a fully adequate investigation are immense. While the data at our disposal are unusually extensive, they are not ideal, and they cannot provide a definitive estimate of the degree of cross-cultural similarity of structures. As will be seen, however, our results do suggest that the similarities may be substantial, and in so doing they suggest that further investigations along this line seem promising.

2. DATA

The data analyzed here come from representative national surveys of the non-institutionalized adult populations in the following nine countries: USA, France, Great Britain, Germany, Italy, Netherlands, Belgium, Denmark, and Ireland. The American data are those of Andrews and Withey (1976) and were collected in May 1972.[4] The European data come from a series of parallel surveys conducted by the European Economic Community and were collected in each of the EEC countries in May 1976.[5] The American survey includes 1297 respondents; each of the eight national European surveys includes approximately 1000 respondents (range 923 to 1047). All of the surveys were conducted by personal interviews using professional field staffs and methods such as to suggest that the data include no unusual quality problems. Interviews were conducted in the native language of the respondents.

In the American survey more than 60 questions asking for evaluations of various life concerns were answered by the respondents. The European data include fifteen such items, of which 11 are reasonably similar to those in the American data. Exhibit 1 presents the exact wording of these 11 items as presented to the American respondents and to English-speaking European respondents.

The American respondents recorded their feelings about these life concerns along a seven-point scale that ranged from "Delighted" to "Terrible," or in one of several off-scale categories: "Neutral (neither satisfied nor dissatisfied)," "I never thought about it," or "Does not apply to me."[6] The European ratings were along an eleven-point scale of satisfaction that ranged from "Completely dissatisfied" to "Very satisfied." While the 7-point Delighted-Terrible and 11-point Satisfaction scales are not identical, previous research suggests that the substantive differences between them are likely to be rather small and that both offer effective means of measuring evaluations of life concerns (see Andrews & Withey, 1976, Chapters 3 and 6).

3. ANALYSIS METHODS

Our interests required the performance of two distinct analytic tasks: (a) identification of the structure of well-being assessments in each country and (b) determination of the similarities among these structures.

The structures were identified using Smallest Space Analysis,[7] one of the several forms of non-metric multidimensional scaling (Guttman, 1968; Shepard, Romney, and Nerlove, 1972). Within each country, associations (product-moment r's) between each pair of the

EXHIBIT 1. Items Used to Assess Evaluations of Life Concerns in American and European Surveys

Reference	American wording	European wording
(Lead in)	In the next section of this interview we want to find out how you feel about parts of your life and life in this country as you see it. Please tell me the feelings you have now--taking into account what has happened in the last year and what you expect in the near future.	Now I would like you to indicate on this scale to what extent you are satisfied with your present situation in the following respects . . .
house	Your house/apartment	The house, flat or apartment where you live
neigh	This particular neighborhood as a place to live	The part of the town or village you live in
income	The income you (and your family) have	The income of you and your family
std lvg	Your standard of living--the things you have like housing, car, furniture, recreation and the like	Your standard of living; the things you have like furniture, household equipment, and so on
job	Your job	Your present work - in your job or as a housewife
spare time	The way you spend your spare time, your non-working activities	The way you spend your spare time
transpt	The way you can get around to work, schools, shopping, etc.	Your means of transport - the way you can get to work, schools, shopping, etc.
health	Your own health and physical condition	Your present state of health
time	The amount of time you have for doing the things you want to do	The amount of time you have for doing the things you want to do
treated	The way other people treat you	The respect people give you
get on w peop	How you get on with other people	In general terms, your relations with other people

well-being assessments were determined, and the resulting matrix of intercorrelations was used as input to Smallest Space Analysis. Smallest Space Analysis then iteratively approaches that configuration of points (i.e., of life concern assessments) in multidimensional space which maximizes the similarities of rank orderings of the distances between the pairs of points and the associations (correlations) between the respective life concern assessments. Thus, assessments that show strong positive associations with one another, suggesting that they tap the same life concern or highly related ones, are placed close to one another, and assessments that are statistically independent are placed far apart. Of course, given a large number of life concern assessments, there is no necessity that a perfect consistency can be achieved between the distances of the points in a small-dimensioned space and the sizes of the associations among the assessments; however, several statistics are available for measuring this consistency.[8]

In the present analysis, structures of subjective well-being were identified by using all of the available well-being assessments--more than 60 assessments in the American data and all 15 items in the European data. Although only 11 assessments were similar between the American and European surveys, the placement of these 11 within each national structure could be more accurately determined within the larger set than if associations among only the 11 matched items were used. After several trials it was determined that a three-dimensional space permitted an adequate portrayal of the structures.[9]

The second major analysis task was to determine the similarity between the various national structures, represented by the three-dimensional configurations of 11 items, as extracted from the larger structures. The rigid ("procrustean") approach proposed by Schönemann and Carroll (1970) was used to match the configurations, and then the degree of match was measured by the Lingoes-Schönemann S statistic (Lingoes and Schönemann, 1974).[10]

The technique of matching involves taking one configuration as the "target" and then rotating, moving, and contracting or dilating another configuration so as to get the corresponding points in each configuration to match one another as closely as possible. Note that the right (90°) angles between the axes are kept rigid and that none of these several transformations changes the relative distances among the pairs of points within either of the configurations; the transformations merely serve to remove inconsequential differences in the original locations, orientations, and sizes of the configurations.

The Lingoes-Schönemann S statistic has two characteristics that make it well suited for assessing configurational similarities in our analysis: (a) It is a symmetric statistic--i.e., it has the same value regardless of which configuration is used as the target. (b) It is scale-invariant--i.e., the value of the statistic does not depend on the "size" of the configurations. These two characteristics are particularly desirable in the present analysis, where our desire to measure the similarity among all possible pairs of nine configurations make it impossible to use the same target for all comparisons.

Since the S statistic is not yet well known, it may be helpful to comment on its interpretation. Lingoes and Schönemann (1974, page 426) note that $S^{1/2}$ is the matrix analogue of a coefficient of alienation (= $(1-r^2)^{1/2}$). Thus low S values imply high similarity (low alienation) and high values imply low similarity. For example, an $S^{1/2} = 1.0$ implies a zero product-moment correlation between the dimensional locations of the points in the two configurations, and an $S^{1/2} = 0.0$ implies a perfect match (product-moment r = 1.00). As will be seen in the following section, values of $S^{1/2}$ of .5, .6, or .7 were typical for the configurations matched here, and these values of $S^{1/2}$ correspond to product-moment correlations between the dimensional locations of .87, .80, and .71, respectively.

So far as we are aware, there have not been, as yet, any statistical tests developed for the S statistic,[11] nor any explorations of how S is affected by various types of measurement errors in the variables that define the configurations. With respect to tests of S, it seems likely that the Schönemann-Carroll transformations, which take advantage of whatever matchings that exist between two configurations, would act to decrease the expected value of S (e.g., pairs of perfectly random configurations would probably show mean values of S below the theoretical S value of 1.00). On the other hand, the impact of measurement errors on the variables probably acts to increase the value of S (e.g., two identical latent configurations, each represented by data containing different measurement errors, would probably not generate the theoretical S = 0). It is virtually certain that both of these effects have influenced the S values reported in the next section,[12] but the extent to which the two effects may have canceled each other is unknown.

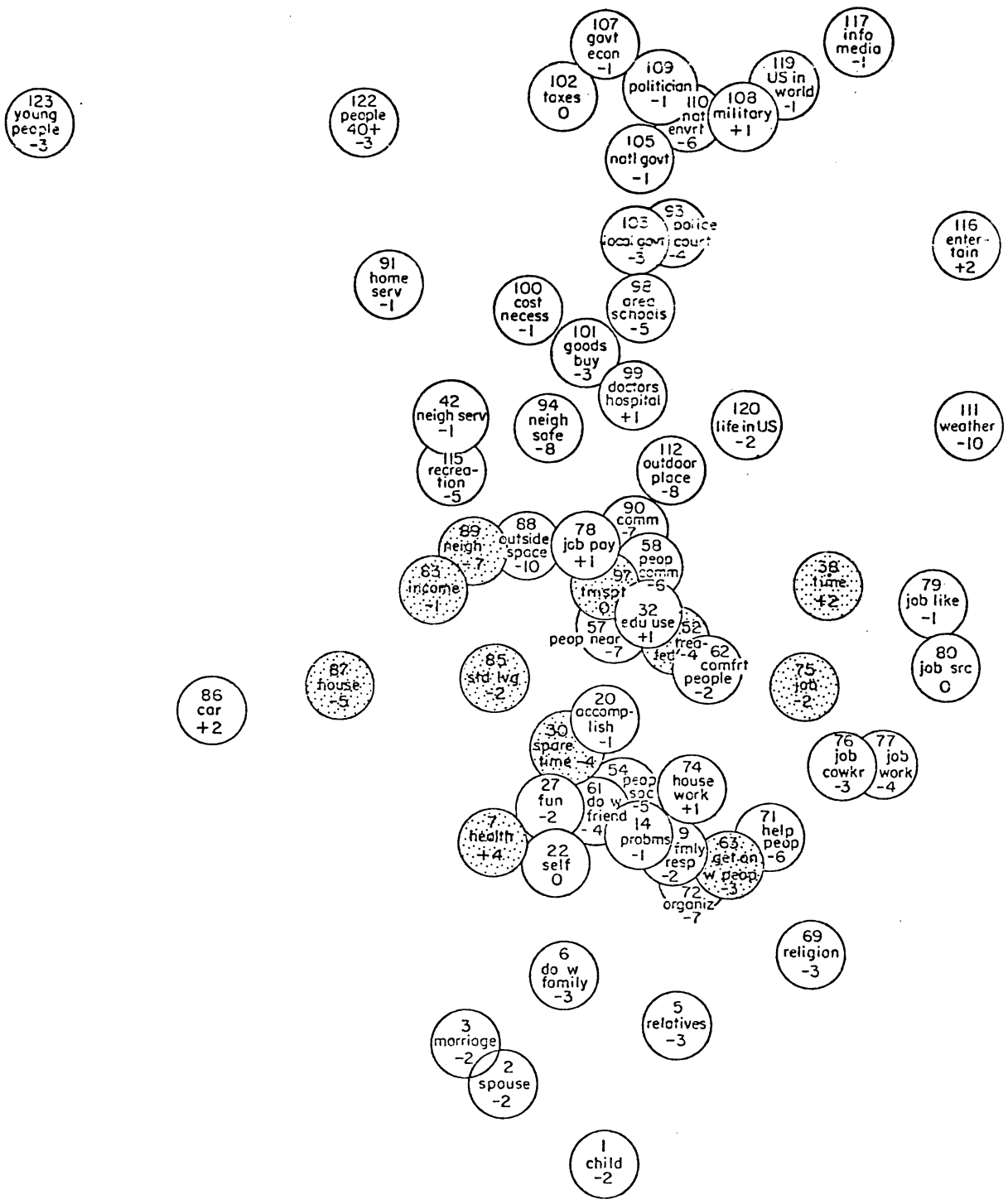
4. RESULTS

It will be most convenient to begin the presentation of results with the configuration for the 60+ life concern evaluations by the American respondents and to note how the 11 items that are similar to those in the European data fit within this larger structure. Following this, we shall examine similarities in the structures for all possible pairs of countries. Finally, we present plots of the structures in selected European nations and of a derived combined configuration for all eight European nations, and compare these structures to that for the USA.

Structure for USA

Exhibit 2 shows the three-dimensional structure for evaluations of 63 life concerns by American respondents and indicates the 11 items from this set that are similar to those used in the European surveys. Several things are worth noting.

EXHIBIT 2. Three-dimensional Structure of Evaluations of 63 Life Concern Items by American Respondents



Notes: Stippled items are those for which similar items exist in European surveys. Signed numbers indicate position on the third dimension. Data source: 1297 respondents to 1972 American national survey. Based on Exhibit 2.4 of Andrews and Withey (1976). For exact wording of all items see Exhibit 2.1 of Andrews and Withey (1976); wording of stippled items appears in Exhibit 1 of the present paper.

(a) One dimension, shown vertically in the exhibit, seems to array items according to the psychological immediacy of the life concern. The dimension ranges from items tapping family concerns (near the bottom of the exhibit), through items tapping concerns about one's relations with the immediate external environment--job, neighborhood, relations with other people, etc. (in the middle of the exhibit), to items tapping concerns about the larger society--national government, mass media, etc. (near the top).[13]

(b) Items that, on the basis of their content, would seem to tap the same life concern do in fact tend to cluster together and thereby serve to locate the nature and approximate position of the underlying concern. For example, note the cluster of job items at the right side of the exhibit, the cluster of family items at the bottom, the cluster of government items at the top, and many others.[14]

(c) The 11 items that are similar to items in the European data (stippled in Exhibit 2) represent a rather limited middle segment of the total structure identified for American respondents. The European data contain no items that are similar to items at the extremes of the vertical (the psychological immediacy) dimension: There are no items at all that tap concerns about marriage or family, and those that tap more remote societal concerns were substantially different from those used in the American survey. Thus what appears to be a major dimension of the American structure will be, of necessity, rather attenuated in the structural matches that follow.

(d) Despite the restricted structural differentiation of the 11 items that are similar to those in the European surveys, a careful examination shows some interesting locational differences. We shall pause to detail them here so that later we can compare them with the European structures. With respect to the first two dimensions of the exhibit (the vertical and horizontal dimensions), one can see that the more personally immediate items--assessments of health, of one's relations with other people, of how one spends one's spare time, and of the amount of time available--are in the lower or right-hand portions of the structure, while items assessing more psychologically remote economic or physical concerns--housing, neighborhood, income, standard of living, and transportation--are in the upper-left portion of the structure. On the third dimension (which runs from "in front of" to "in back of" the plane of the exhibit), the housing and neighborhood items are well "back," the income and standard of living items, the two items tapping relations with other people, and the spare time item are modestly "back," and the health item is somewhat in "front."

Similarity Among Nine Countries

Having examined the structure of subjective well-being assessments in some detail as derived for American respondents, we can now ask how similar it is to comparable structures for respondents in eight European countries. We can also ask how similar the European structures are to one another. Some initial answers appear in Exhibit 3, which presents values of $S^{1/2}$ for all possible comparisons among the nine countries.[15] Also shown in Exhibit 3 is the similarity of each national structure to a derived structure which represents the single best-fit approximation to the eight individual European structures.

EXHIBIT 3. Degree of Dissimilarity Between Structures of Life Concern Assessments in Nine Countries

	USA	FRA	GB	GER	ITA	NLD	BEL	DEN	IRE
USA	---								
FRA	.77	---							
GB	.78	.44	---						
GER	.71	.74	.66	---					
ITA	.70	.66	.67	.72	---				
NLD	.64	.70	.64	.72	.72	---			
BEL	.75	.52	.56	.73	.81	.63	---		
DEN	.77	.68	.63	.73	.75	.46	.55	---	
IRE	.65	.69	.56	.54	.56	.58	.76	.72	---
***	.65	.61	.57	.67	.68	.62	.64	.63	.61

*** European centroid

Notes: The measure of dissimilarity is $S^{1/2}$, a matrix alienation coefficient (Lingoes and Schönemann, 1974). Low values of $S^{1/2}$ indicate high configurational similarity.

The left-most column of the exhibit shows how the USA structure (of 11 items, as contained within the larger set of 63 items shown in Exhibit 2) matches each of the European national structures (of 11 items as contained within their own larger sets of 15). One can see that the coefficients vary only modestly--from .64 to .78. This suggests that Americans' structure of well-being perceptions is about as similar to the structure of one European country as it is to another. Within the limited range of the differences, however, the American structure is most similar to that of The Netherlands, closely followed by Ireland, and least similar to the structures in Great Britain, France, and Denmark.

The fact that the British structure is least similar to the American has interesting implications: It suggests that the cross-national differences we observe do not reflect artifacts of translation, for the wording of the British and American items was closely similar (in some cases identical), yet the differences between the American and British structures are greater than those between the American pattern and that

resulting from questions posed in German, French, Dutch, Danish, or Italian.

Probably more important than these modest differences, however, is the absolute level of the coefficients in the left-most column of the exhibit. With values approximating .7 (which, as noted in Section 3 of this paper, correspond to product moment r's of about .7), the data suggest a rather substantial configurational similarity between structures of well-being assessments in the United States and these European countries.

The value of .65 shown for the match between the American configuration and the European centroid configuration is also of interest. This figure suggests that the European average is somewhat closer to the USA structure than are most of the individual European countries. Thus while the European average structure is certainly not identical to the American structure, as the individual European national structures deviate away from their own average, they also tend to deviate away from the American structure rather than toward it. Or in still other terms, the American structure has (slightly) more in common with Europe-as-a-whole than with most of the individual European structures.

Exhibit 3 also provides interesting results on the similarities among the various European structures themselves. Here the coefficients vary from .44, for Great Britain and France (which are most similar to one another), to .81, for Belgium and Italy (which are least similar). Furthermore, if one computes some averages based on the data in Exhibit 3 one finds that of all the individual national structures, the British structure is most typical of the European structures (mean $S^{1/2}$ = .59) and the Italian structure is most distinctive (mean $S^{1/2}$ = .70). The same results can be seen in Exhibit 3 by comparing the individual European structures to the European centroid.

To summarize these various findings from Exhibit 3 we can observe that: (a) there seems to be a basic similarity in structures among all nine of these western societies; (b) within this basic similarity the European structures are distinct from the American structure; (c) even within Europe there is modest heterogeneity; and (d) if one averages out the differences among the individual European structures, the result is a structure that is closer to the American structure than are most of the individual European structures.[16]

Structures for EEC Countries

What are the European structures? Lack of space precludes a presentation of each one, but Exhibits 4 and 5 present the structures for The Netherlands and Great Britain, respectively.[17] The Dutch structure was selected because it is the individual structure most similar to the American one; and the British structure because, while still basically similar, it matches the USA least well. As presented in Exhibits 4 and 5, the Dutch and British structures are oriented to agree with the presentation of the American structure in Exhibit 2.

EXHIBIT 4. Three-dimensional Structure of Evaluations of 11 Life Concern Items by Dutch respondents

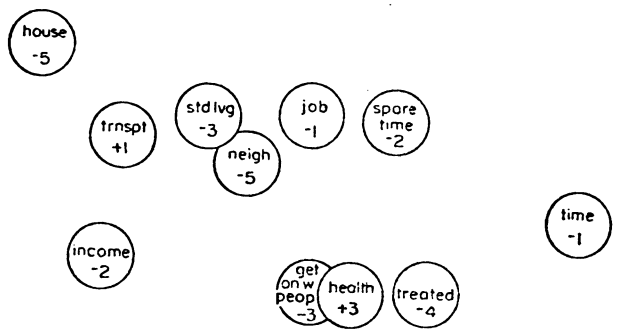
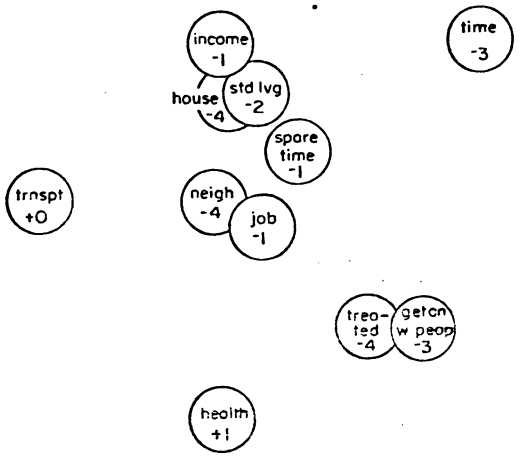


EXHIBIT 5. Three-dimensional Structure of Evaluations of 11 Life Concern Items by British Respondents



In the case of both the Dutch and British structures one can see the same basic pattern among these 11 items that was identified previously in our discussion of the American structure. Note that all the personally immediate items (health, relations with other people, spare time activities, and amount of time available) are located in the lower or right-hand portions of the structures, while the more psychologically remote economic or physical concerns fall in the upper left portions. Note also that, comparable with the American structure, housing and neighborhood are both well "back" on the third dimension, that the two items that tap relations with other people, the income and standard of living items, and the spare time item are modestly "back," and that the

health item is somewhat in "front" of the plane of the exhibit.[18]

Another view of the similarities and differences between the American and European structures is presented in Exhibit 6. Plotted there is the United States structure (reproduced from Exhibit 2 and shown by the U's) and also the best-fit average European structure (the European centroid configuration that was discussed previously and for which similarity measures were presented in Exhibit 3--shown by the E's).

The basic similarity of the two configurations in the first two dimensions is indicated by the fact that most of the linkage lines are relatively short, and in the third dimension by the close similarity in values of the signed numbers.

With this exhibit it is easy to identify the modest differences that do exist. In the European centroid configurations, compared to the United States structure: the two items tapping relations with other people are much closer together; the job item is much closer to assessments of income and standard of living; and the transportation item moves out to a less central position while income, housing, and neighborhood evaluations move into more central positions. While one might speculate about the inter-cultural causes of these discrepancies, we feel such speculation is best avoided for the present. These differences may be at least partly real, but they are almost certainly at least partly the result of various methodological artifacts.[18]

We believe that what is most important is the basic similarity of the structures.

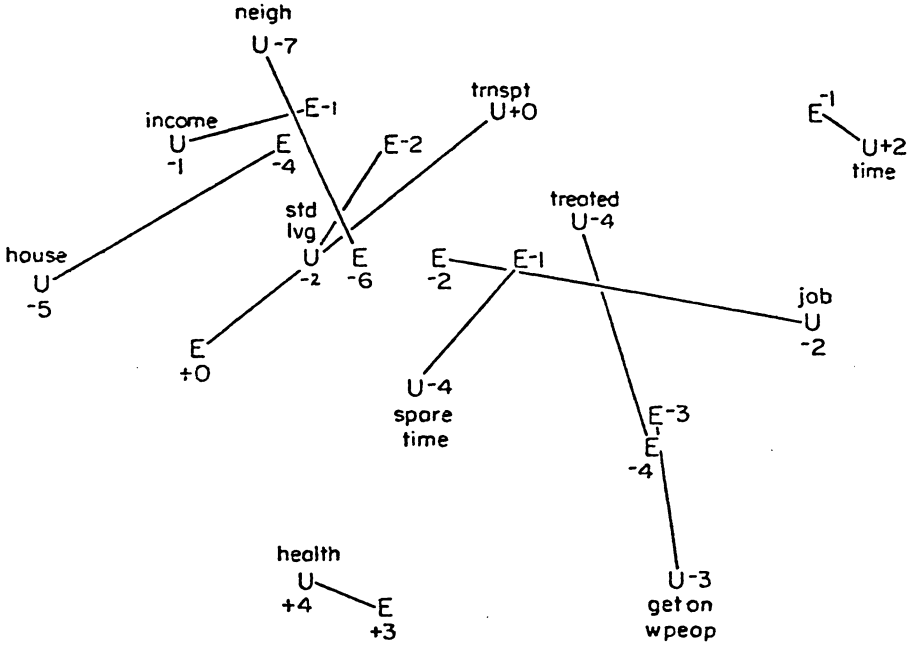
5. COMMENTS AND CONCLUSIONS

The basic similarities between the structures of well-being assessments across the nine western societies examined here is, we believe, an important finding for the social indicators movement. It suggests that it is feasible to do cross-national comparative research--at least in these countries--on the topic of perceived well-being and that meaningful comparable results can be expected from the use of standardized survey instruments and methods. Only if people in different societies think about well-being in basically similar ways would this be the case, and the initial explorations reported here suggest that in fact they do.

The limited nature of the present explorations, and hence the tentativeness of the conclusions, must be recognized. The number of well-being assessments that were reasonably similar across the national surveys at our disposal was only eleven, and it happens that these eleven items represent only a portion of what previous research suggests is the larger structure. In addition, this analysis is limited by certain methodological factors, including differences in the times at which the American and European surveys were conducted (1972 versus 1976), differences in the response scales used by the two sets of respondents (7-point Delighted-Terrible versus 11-point Satisfaction) and differences in the wording and linguistic translations of the

EXHIBIT 6. Match Between Three-dimensional Structures of Evaluations of Life Concern Items by American and European Respondents

Key: U = location in United States structure
E = location in European centroid structure



items. Some of these differences are endemic to any cross-national research and will always compete with the hypothesis that observed differences are attributable to cultural effects, but even within the limits of what is feasible in current cross-cultural research, one could design data better suited to address the issue of structural similarity.

Besides the restrictive nature of the data at our disposal, the definitiveness of the results is limited by the lack of statistics for testing the significance of similarities between configurations and the limitations of knowledge regarding how various types of measurement errors affect measures of configurational similarity.

Assuming that these data and statistical limitations may some day be removed, we would propose some promising extensions of this line of research. (a) Of course, we would wish to extend the descriptive data about similarity of structures beyond the nine western countries examined so far. Would other western countries show similar patterns? What about well-being perceptions in non-western countries? (b) To the extent that significant differences in structures of well-being assessments were identified, one would want to move beyond the descriptive phase and begin to ask what accounts for the structural differences and what impact they have on the behavior of people, governments, etc. Even within the range of the modest differences noted among the nine western societies investigated here, there are hints that similarity varies directly with geographical contiguity, with the comparability of the socioeconomic systems, and/or with the general level of well-being.[20] Any conclusion along these lines, however, must be extremely tenuous with the present data and would have to be checked against results for a wider range of societies. (c) We have in this paper identified national structures (plus one regional structure--the European centroid configuration). While national structures are conceptually convenient and have an obvious interest, it is possible that other groups of persons should be considered. One can imagine cross-national groupings based on characteristics such as age, sex, occupation, socioeconomic status, cultural group, language, and others. (One can also imagine performing the analysis on certain sub-divisions within a given national grouping, but the research on this topic to date suggests that structural differences are modest.[21])

FOOTNOTES

[1]Prepared for presentation at the 1977 Annual Meeting of the American Statistical Association, Chicago, August 1977. We are grateful to Kai Hildebrandt for his skillful processing of data for this paper and for many useful suggestions regarding the analysis. Ed Schneider and James Lingoes also provided helpful advice.

[2]See, for example, Abrams (1974); Allardt (1975); Andrews and Withey (1976); Campbell, Converse, and Rodgers (1976); Development Academy of the Philippines (1975); Hall (1976); Inglehart (1977); Rabier (1974); and Riffault and Rabier (1977).

[3]Structural similarity does not, of course, imply that all societies will be similarly satisfied--either in general or with respect to specific life concerns; rather, it means that the relationships among the well-being assessments will be similar.

[4]Collection of these American data was supported by grant GS3322 from the National Science Foundation. These data, together with four other sets of survey data on perceptions of well-being collected under the direction of Andrews and Withey, are available from the Social Science Archive of The Institute for Social Research, The University of Michigan, Ann Arbor, Michigan and also from the Inter-university Consortium for Political and Social Research.

[5]These European data are extracted from the May 1976 Euro-Barometer, a series of national surveys conducted semi-annually in the EEC countries and coordinated by the Commission of the European Community. For more details on these surveys and a report of results from earlier Euro-Barometers, see Inglehart (1977), Rabier (1974), and Riffault and Rabier (1977). These and other data from the series are available from the Belgian Archives for the Social Sciences, Catholic University, Louvain, and also from the Inter-university Consortium for Political and Social Research.

[6]The off-scale categories were rarely used (with obvious exceptions, such as inquiries about "job"), and were treated as missing data.

[7]The technique is implemented in a computer program called MINISSA (Roskam and Lingoes, 1970; Lingoes and Roskam, 1973; Lingoes, Guttman, and Roskam, 1977). Input to MINISSA was a matrix of Pearson correlation coefficients (computed with pairwise deletion of missing data).

[8]We have, above, likened the identification of psychological structures to "cognitive mapping." This analogy is legitimate: If one submits a matrix of distances between geographic points (e.g., cities) to Smallest Space Analysis, it will produce an acceptable geographic map of the region involved.

[9]The alienation coefficient, a measure of the consistency between the interpoint distances in the multidimensional space and the intercorrelations among the life concern assessments, ranged from .10 to .13 for the eight European countries when 15 items were arrayed in three-dimensional space, and was .19 for USA when more than 60 items were arrayed in three-dimensional space. Comparable figures for two-dimensional space were .18-.21 for the European countries and .26 for USA. When only the 11 items that are similar in the USA and European data were arrayed in three dimensions, the coefficient of alienation for the USA data was .10.

[10]Two computer programs were used to accomplish these tasks: PINDIS (Lingoes and Borg, 1976; Lingoes, Guttman, and Roskam, 1977), and SPACES (Computer Support Group of the Center for Political Studies, 1976).

[11]Neel, Rothhammer, and Lingoes (1974) report a Monte Carlo exploration of the stability of S in one application but do not provide statistical tests which are of general applicability.

[12]From previous analyses (Andrews and Withey, 1976, Chapter 6), we can estimate that the American data used in this paper have validity of about .7, reliability of about .8, and include about 10 percent correlated measurement error and about 40 percent uncorrelated measurement error. A roughly similar composition is expected to characterize the European data.

[13]The two other dimensions of the space, while needed to locate items in correct relative position to one another, do not seem to show conceptually meaningful progressions. While such progressions are interesting if found, there is no necessity that they occur, and no requirement that one "interpret" the dimensions of a structure. (Note that the same applies to the 2--or 3--dimensions of geographic or celestial maps.)

[14]Andrews and Withey (1976, Chapter 3) identify 12 clusters among these items.

[15]Values of the square root of S (i.e., of $S^{1/2}$) rather than of S are presented because it is this statistic that Lingoes and Schonemann (1974) propose as the matrix analogue of a coefficient of alienation (and because these values are produced by the PINDIS and SPACES computer programs used for the present analysis).

[16]In an exploration going one step beyond the similarity analyses reported in this subsection, we considered the possible effects of differential weighting of the dimensions of the configurations. (This is another capability of the PINDIS and SPACES computer programs referenced previously.) While differential weighting made it possible to more closely match most of the configurations, the differences were not large and the basic pattern of results just described for Exhibit 3 was maintained.

[17]Configurations for the other six European countries will be provided upon request.

[18]Campbell, Converse, and Rodgers (1976, pp. 74-75) report a matching of American and British structures based on different and somewhat more limited data than those used for Exhibits 2 and 5. While some of the details of their matching differ from what we find here, their general conclusion--that "the correspondence is fairly close"--clearly agrees with ours. Levy (1976) has also reported a matching of well-being structures derived from American and Israeli respondents. Here, also, a conclusion of there being a substantial match was also put forward, though a numerical assessment of the degree of fit was not made.

[19]These include differences in item wordings (particularly for the job item), in the set of other items that were present when the original structures were determined, and in the error compositions of the measures.

[20]For example, we observed a correlation of about .3 between the similarity of structures of perceived well-being (shown in Exhibit 3) and the differences between the countries in mean satisfaction with "life in general." Given a more heterogeneous set of countries, this relationship might appear stronger and, if so, might be attributed to the operation of Maslovian principles.

[21]Andrews and Withey (1976, Chapter 2) and Campbell, Converse, and Rogers (1976, Chapter 3) both report explorations of differences in such perceptual structures among subgroups of the American population. Both sets of investigators, using entirely independent sets of data, came to the same general conclusion: that while modest differences appeared among structures identified for the subgroups, the basic features of the structure identified at the national level remained evident. Andrews and Withey (1976, Chapter 4) also showed that the same prediction equation was about equally effective for a large number of different subgroups for predicting feelings about general well-being on the basis of evaluations of life concerns.

REFERENCES

Abrams, M. Subjective social indicators. Social Trends, No. 4. London, Her Majesty's Stationary Office, 1973.

Allardt, E. Dimensions of welfare in a comparative Scandinavian study. Research Reports No. 9. Helsinki, Research Group for Comparative Sociology, University of Helsinki, 1975.

Andrews, F. M. and Withey, S. B. Social Indicators of Well-being: Americans' Perceptions of Life Quality. New York, Plenum, 1976.

Campbell, A., Converse, P. E., and Rodgers, W. L. The Quality of American Life: Perceptions, Evaluations, and Satisfactions. New York, Russell Sage Foundation, 1976.

Computer Support Group of the Center for Political Studies. SPACES: Program Writeup. Ann Arbor, Michigan, Institute for Social Research, The University of Michigan, June 1976.

Development Academy of the Philippines. Measuring the Quality of Life: Philippine Social Indicators. Manila, Development Academy of the Philippines, 1975.

Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, 1968, 33, 469-506.

Hall, J. Subjective measures of quality of life in Britain: 1971 to 1975; Some developments and trends. Social Trends, No. 7. London, Her Majesty's Stationary Office, 1976.

Inglehart, R. F. The Silent Revolution: Changing Values and Political Styles among Western Publics. Princeton, Princeton University Press, 1977.

Levy, S. Use of the mapping sentence for coordinating theory and research: A cross-cultural example. Quality and Quantity, 1976, 10, 117-125.

Lingoes, J. C. and Borg, I. Procrustean individual difference scaling. Journal of Marketing Research, 1976, 13, 406-407.

Lingoes, J.C., Guttman, L., and Roskam, E. E. Geometric Representations of Relational Data. Ann Arbor, Michigan, Mathesis Press, 1977.

Lingoes, J. C. and Roskam, E. A mathematical and empirical study of two multidimensional scaling algorithms. Psychometric Monographs, 1973, 38.

Lingoes, J. C. and Schönemann, P. H. Alternative measures of fit for the Schönemann-Carroll matrix fitting algorithm. Psychometrika, 1974, 39, 423-427.

Neel, J. V., Rothhammer, F., and Lingoes, J. C. The genetic structure of a tribal population, the Yanomamo Indians. X. Agreement between representations of village distances based on different sets of characteristics. American Journal of Human Genetics, 1974, 26, 281-303.

Rabier, J. R. Satisfaction et Insatisfaction Quant aux Conditions de Vie dans les Pays Membres de la Communaute Europeene. Brussels, Commission of the European Communities, 1974.

Riffault, H. and Rabier, J. R. The Perception of Poverty in Europe. Brussels, Commission of the European Communities, 1977.

Roskam, E. and Lingoes, J. C. MINISSA-I: A FORTRAN IV(G) program for the smallest space analysis of square symmetric matrices. Behavioral Science, 1970, 15, 204-205.

Schönemann, P. H. and Carroll, R. M. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, 1970, 35, 245-255.

Shepard, R. N., Romney, A. K., and Nerlove, S. B. Multidimensional Scaling. New York, Seminar Press, 1972.

INNOVATIONS IN THE SAMPLE DESIGN FOR THE SURVEY OF INCOME AND EDUCATION

Donald M. Luery and Gary M. Shapiro, U.S. Bureau of the Census

I. INTRODUCTION

This paper discusses in detail some of the more interesting aspects of the sample design of the Survey of Income and Education (SIE) and should be of prime interest to people engaged in designing complex surveys. Some other aspects of the sample design for this survey are covered in detail in Boisen [1] and are briefly discussed in this memorandum.

Only 6-9 months' time was available to decide on and execute the sample design for this survey. Optimality criteria were generally applied in determining the sample design, but the application was generally imperfect.

The SIE was designed to meet three major objectives. Title I of the Elementary and Secondary Education Act of 1965 [10] provided for the annual distribution of $2,000,000,000 to local school districts, with the intent that school districts servicing low income areas should receive relatively more money than school districts servicing high income areas. One provision of the Educational Amendments of 1974 [11] to this Act states that the Secretary of Commerce shall "expand the current population survey (or make such other survey) in order to furnish current data for each State with respect to the total number of school age children in each State to be counted for purposes of Section 103(c)(1)(A) of Title I of the Elementary and Secondary Act of 1965." Thus, the prime objective for the SIE was its use in conjunction with the existent Current Population Survey[1] (CPS) to produce estimates of children, age 5-17, in poverty families with coefficients of variation of 10 percent or better by State.

Another section of the same law dealt with questions of bilingual education and required the Office of Education in the Department of Health, Education and Welfare (HEW) to issue a report to Congress including among other things, "...a national assessment of the educational needs of children and other persons with limited English-speaking ability ..." (PL93-380). This leads to a secondary purpose for SIE of providing estimates of persons with limited English-speaking ability by State. The questions relating to language ability were to be asked only on the SIE questionnaire, not on the CPS questionnaire, and hence, the language ability tabulations were to be based only on the SIE.

The tertiary objective of the SIE was to provide cross-tabulations involving poverty and other items from SIE by itself that were of interest to analysts in the Department of Health, Education and Welfare. The reason that CPS was not to be used for these tabulations was that the SIE questionnaire contained additional questions on food stamp recipiency, housing costs for homeowners and renters, estimated cash receipts, education, disability, and health insurance coverage.

Initially, SIE was intended to have a designated sample size of 200,000 housing units. The methods used in deciding how this was to be allocated by State, consistent with the three objectives discussed above, are discussed in Section II. For budgetary reasons, it was decided after the basic sample was selected that a sample reduction to about 190,000 designated units, resulting in about 151,000 interviewed households, was needed. This required reallocation of sample by State is discussed briefly in Section II.

The SIE was designed completely independently of the CPS on a State-by-State basis, except that the primary sampling unit (PSU) definitions were the same. In most States, primary sampling units consisting of SMSA's or groups of counties and independent cities, were divided into strata according to estimates based on 1970 census data of the proportion of persons who were children age 5-17, living in poverty families. PSU's in a State that were large enough to provide at least 80 sample housing units formed a stratum by themselves and came into sample with certainty. In nine States (Conn., Del., D.C., Hawaii, Md., Mass., N.H., R.I., and Vt.) every PSU was selected with certainty. In the remaining States, from one to ten non-self-representing strata with three or more PSU's were formed in each State. Two sample PSU's were selected with replacement from each stratum using the Durbin-Sampford rejective method. See Durbin [5] and Sampford [7].

The major frame for sampling housing units from a selected PSU was the list of units enumerated in the 20 percent sample of the 1970 census. The 20 percent sample was used instead of the full census file because of the information on income and poverty available from it. Two methods of selection were employed in the selection from the census file. For the first method, some enumeration districts (ED's) were selected and a sample of approximately three housing units was selected from each ED. (An ED was the assignment given to a single interviewer in the 1970 census. On the average, an ED contains approximately 350 housing units). For the second method, a direct selection of housing units was taken without the intervening step of selecting ED's. These two methods of selection are more fully described in section IV of this paper. In order to attempt full coverage of housing units, a systematic sample from four additional frames was selected: (1) special places, (2) units built since the 1970 census in jurisdictions that issue permits, (3) units built since the 1970 census in jurisdictions that do not issue permits, and (4) mobile homes in parks established since the 1970 census.

Section III of this paper discusses the methods used to decide that noncompact clusters of three housing units should be used for most States.

Section IV discusses why the Durbin method was used for most stages of selection, how it was applied, the difficulties caused by the required reduction in sample size, and some advantages and disadvantages of the Durbin method.

II. ALLOCATION OF SAMPLE BY STATE

Sample was allocated to each State in accordance with the three primary objectives of the survey as stated above and the amount of money available for the survey. Most of the credit for the allocation scheme which is described should go to Mr. Wray Smith in the Office of the Secretary, Department of Health, Education, and Welfare. The

authors assume full responsibility, however, for any errors in this paper and any problems of logic with the allocation scheme.

The sample was not allocated in one stage but rather in three stages, one stage for each primary objective. The vast majority of the sample was allocated to satisfy the first objective of producing estimates of children, age 5-17, in poverty families. However, this was done in the first stage, so that the allocation decisions for the other two objectives could take advantage of the large sample intended to satisfy the first objective. Had sample been allocated in one stage or in a different order to meet the three objectives, there would have been substantial differences in sample size for some States.

We began with the sample present in the CPS including the supplementation to CPS begun in July 1975. (See Dippo [4] for details on this supplementation.) The sample totals are given by State in column (2) of table 1. We determined the additional sample needed for each State to achieve an expected 9.6 percent coefficient of variation on the estimated children, aged 5-17, in poverty families in the State. The choice of 9.6 percent was somewhat arbitrary. The criteria had to be a coefficient of variation less than or equal to 10.0 percent; 9.6 percent was an affordable criteria and brought a little bit of safety for achieving a true 10.0 percent coefficient of variation in each State. A number of assumptions were needed to determine the sample sizes. Perhaps the most important was an estimate of the number of children in poverty families. Rather standard methodology was used, however, so no description will be given here. Details are given in appendix A of Boisen [1] and there is some related discussion in section III of this paper. The supplemental sample sizes to meet this objective are given by State in column (3) of table 1.

Next we allocated about 36,000 sample households to the States to improve the estimates of persons with difficulty speaking English. These estimates were to be made from the SIE sample only. The 36,000 figure corresponded roughly with the amount of money being contributed to the total survey effort by the part of HEW interested in these estimates. This additional sample was allocated in order to bring the total allocation closer to optimal allocation, according to the standard optimum allocation formula, for a national estimate of persons with difficulty speaking English. The second objective is, of course, concerned with State, not national estimates. However, there was no requirement for equal reliability for each State and, in fact, it was felt that States with a relatively serious problems of persons with difficulty speaking English needed greater reliability in their estimates. Optimally allocating a sample for a national estimate is one way of achieving this. At the same time, it was desired that all States have a reasonably large sample size for the planned analysis and 2,000 was selected as a minimal supplementary sample size per State for this purpose.

Finally, we allocated sample households to the States to improve estimates of children in poverty based on the SIE sample only, without benefit of the CPS sample. The criteria was a 9.9 percent CV on State estimates. The third objective does not relate specifically to total children in poverty and the choice of 9.9 percent CV is completely arbitrary other than it being consistent with the total sample size that could be afforded. It was felt, however, that this allocation would well serve the third objective. Sample sizes are given in column (5) of table 1.

All of the above relates to the original allocation before the budget-imposed reduction. The reallocation necessitated by the reduction was accomplished through similar procedures. Instead of a 9.6 percent CV criteria for the first allocation, a 9.8 percent CV criteria was used; this reduced the sample allocated in this stage from 157,000 to 148,000. The procedure and number of sample cases for the second stage of allocation was completely unchanged. In the third stage of allocation the criteria was 10.4 percent CV instead of 9.9 percent CV; this reduced the sample allocated in this stage from 12,000 to 6,500. The final supplementary sample sizes after reduction are given in column (10) of table 1.

Note that all sample sizes given are originally intended expected sample sizes. The figures were used to determine sampling rates. Application of these sampling rates did not yield the exact figures given in column (10).

III. DETERMINATION OF NONCOMPACT CLUSTERS OF THREE HOUSING UNITS

We started with the assumption that we would generally select a sample of enumeration districts (ED's) from the sample of PSU's and that only a single cluster of housing units (or a single special place hit) would usually be selected from each ED. In order to objectively determine optimal cluster size and to determine whether clusters of housing units should be compact or dispersed throughout an ED (noncompact), several cost figures and intraclass correlations were necessary.

Comparisons for different compact and noncompact cluster sizes were based on estimated design effects; that is, estimates of the increase in variance because cluster sampling of households instead of a simple random sample of persons was used. The characteristic of interest was school-aged children in poverty families. In all the calculations made, we assumed a simple random sample of ED's from the State and, for noncompact clusters, a simple random sample of housing units from ED's. In fact, of course, ED's were not generally selected directly from a State and a systematic sample of housing units was selected for noncompact clusters.

The formula used for the design effect for compact clusters, which measures the increase in variance expected from selecting compact clusters of households as compared to selecting a simple random sample of persons, was[2]

$$(\hat{V}_L^2/\hat{V}_L^2)\ (\hat{V}_K^2/\hat{V}_K^2)\left[1+\delta_{\bar{P}}(\bar{P}-1)\right] \qquad (1)$$

The formula used for the design effect for noncompact clusters of housing units versus a simple random sample of persons was[3]

$$(\hat{V}_L^2/\hat{V}_L^2)\ (\hat{V}_K^2/\hat{V}_K^2)\left[1+\delta_{\bar{\bar{K}}}(\bar{\bar{K}}-1)\right]\left[1+\delta_{\bar{\bar{N}}}(\bar{\bar{N}}-1)\right] \qquad (2)$$

TABLE 1.—Sample Sizes (Housing Units) by State and by Stages of Allocation, and Coefficients of Variation for Important Estimates

STATE	CPS Sample Size (including Supplementation)	Supplement for Children in Poverty Estimate	Supplement for Bilingual Estimates	Supplement for Ests. Based on Sample Excluding CPS	Total Supplementary Sample Size (3)+(4)+(5)	CV's for Children in Poverty Est. Based on Complete Sample	CV's for Children in Poverty Based on Sample Excluding CPS	CV's for Persons with Difficulty Speaking English	Total Supplementary Sample Size After Reduction
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
TOTAL	68790	157253	35969	11906	205127				190243
Maine	900	2512	0	374	2886	0.091	0.099	0.088	2747
New Hampshire	740	5285	0	132	5417	0.095	0.099	0.057	5252
Vermont	670	3413	0	161	3574	0.094	0.099	0.082	3370
Massachusetts	1450	4251	233	420	4903	0.091	0.099	0.055	4458
Rhode Island	580	4212	0	99	4311	0.095	0.099	0.054	4032
Connecticut	890	5479	0	229	5707	0.094	0.099	0.046	5175
New York	4680	1057	4374	0	5431	0.068	0.086	0.043	5221
New Jersey	1850	3779	2042	0	5821	0.081	0.089	0.045	5666
Pennsylvania	3070	2920	2778	0	5698	0.077	0.091	0.065	5464
Ohio	2750	3307	2446	0	5754	0.079	0.090	0.075	5502
Indiana	1550	5201	0	744	5946	0.091	0.099	0.084	4794
Illinois	2770	2704	2963	0	5667	0.075	0.088	0.057	5465
Michigan	2370	3330	2427	0	5757	0.079	0.090	0.067	5514
Wisconsin	1070	4783	0	404	5187	0.093	0.099	0.057	3966
Minnesota	1250	4304	0	473	4778	0.092	0.099	0.065	4084
Iowa	950	4720	0	261	4982	0.094	0.099	0.081	4535
Missouri	1540	2596	71	652	3319	0.087	0.099	0.112	3000
North Dakota	980	3198	0	425	3624	0.091	0.099	0.066	4143 *
South Dakota	1120	1806	75	493	2373	0.087	0.099	0.096	2877 *
Nebraska	800	3651	0	238	3889	0.093	0.099	0.082	3603
Kansas	880	3991	0	279	4270	0.093	0.099	0.091	3979
Delaware	540	3555	0	183	3737	0.094	0.099	0.093	2910
Maryland	980	3663	0	461	4124	0.091	0.099	0.090	3160
District of Columbia	550	2206	0	217	2423	0.092	0.099	0.100	2171
Virginia	1230	2663	0	565	3229	0.089	0.099	0.131	2569
West Virginia	840	2445	0	339	2784	0.091	0.099	0.161	2212
North Carolina	1310	1785	314	274	2373	0.086	0.099	0.180	2114
South Carolina	830	1522	588	0	2110	0.085	0.096	0.200	2000
Georgia	1330	1377	654	41	2071	0.084	0.099	0.177	2000
Florida	2320	2179	961	275	3415	0.083	0.099	0.070	3310
Kentucky	910	1965	0	314	2279	0.090	0.099	0.178	2000
Tennessee	1010	1838	152	335	2325	0.088	0.099	0.182	2175
Alabama	1030	1320	846	0	2166	0.080	0.091	0.202	2199 *
Mississippi	810	766	1187	47	2000	0.068	0.077	0.203	2000
Arkansas	870	1603	354	44	2000	0.088	0.098	0.196	2000
Louisiana	1130	1033	1624	0	2657	0.068	0.076	0.087	2165
Oklahoma	960	2177	0	410	2586	0.090	0.099	0.137	2476
Texas	3270	588	4775	0	5363	0.054	0.062	0.039	5182
Montana	1010	3797	0	327	4124	0.093	0.099	0.083	3823
Idaho	900	6054	0	143	6198	0.095	0.099	0.072	5807
Wyoming	720	4959	0	142	5102	0.095	0.099	0.066	4253
Colorado	970	3352	0	385	3737	0.092	0.099	0.059	4130 *
New Mexico	910	1083	2063	0	3146	0.063	0.068	0.034	2580
Arizona	800	2649	0	334	2983	0.091	0.099	0.052	2657
Utah	900	4612	0	362	4974	0.093	0.099	0.076	5057 *
Nevada	650	5323	0	89	5418	0.095	0.099	0.058	4911
Washington	1000	4646	0	328	4974	0.093	0.099	0.075	4339
Oregon	880	4480	0	231	4711	0.094	0.099	0.079	4896 *
California	5690	278	5041	0	5319	0.065	0.086	0.043	5117
Alaska	1030	2395	0	515	2909	0.089	0.099	0.082	3780 *
Hawaii	550	4434	0	162	4596	0.094	0.099	0.050	3401

NOTE: The first 9 columns of this table appeared in Boisen [2]; column (10) appeared in Smith [8].

* The sample sizes are higher for these States after the reduction allocation than from the original allocation because more accurate data on between PSU variances was available and the variances for these States were higher than previously speculated.

The notation and meaning of these terms is as follows:

V_K^2 is the population relvariance between persons for the proportion of poverty children.

$\hat{V}_K^2$ is like V_K^2 except it includes the effect of the number of persons per listing unit.

$(\hat{V}_K^2/V_K^2)$ is the increase in variance due to variation in the number of persons per listing unit. Note that the listing unit is taken to be a compact cluster of housing units for Formula (1) and a single housing unit for Formula (2). Thus, $(\hat{V}_K^2/V_K^2)$ is not precisely the same quantity in (1) as in (2).

V_L^2 is the population relvariance between housing units for the number of poverty children per household.

$\hat{V}_L^2$ is like V_L^2 except it includes the effect of the variation of the number of housing units per ED.

$(\hat{V}_L^2/V_L^2)$ is the increase in variance due to variation in the number of housing units per ED.

$\bar{\bar{K}}$ is the average number of persons per housing units.

$\bar{\bar{N}}$ is the average number of housing units per cluster.

$\bar{\bar{P}}$ is the average number of persons per cluster. $\bar{\bar{P}} = \bar{\bar{K}}\,\bar{\bar{N}}$.

$\delta_{\bar{\bar{P}}}$ is the intraclass correlation between persons within listing units. For noncompact clusters, $\bar{\bar{P}}=\bar{\bar{K}}$.

$1+\delta_{\bar{\bar{P}}}(\bar{\bar{P}}-1)$ is the increase in variance due to sampling listing units instead of persons, assuming no variation in number of housing units per cluster.

$\delta_{\bar{\bar{N}}}$ is the intraclass correlation between housing units within an ED.

$1+\delta_{\bar{\bar{N}}}(\bar{\bar{N}}-1)$ is the increase in variance due to sampling a cluster of $\bar{\bar{N}}$ housing units instead of a single housing unit.

For calculating $\delta_{\bar{\bar{K}}}$ (for a compact cluster of one housing unit), we obtained a special tabulation from the 1970 decennial census giving, for each family size, the distribution of families with zero children in poverty, one child in poverty, two children in poverty, etc. From this we were able to calculate directly a within family-size group relvariance W^2 and a between family-size group relvariance B^2. From these relvariance estimates, we calculated:

$$\delta_{\bar{\bar{K}}} = \frac{\frac{N-1}{N}B^2 - \frac{W^2}{\bar{\bar{K}}}}{\frac{N-1}{N}B^2 + \frac{(\bar{\bar{K}}-1)W^2}{\bar{\bar{K}}}} \qquad (3)$$

where N is the number of housing units in the U.S., hence $\frac{N-1}{N} \doteq 1$.

We needed $\delta_{\bar{\bar{P}}}$ for values of $\bar{\bar{P}}$ other than $\bar{\bar{P}}=\bar{\bar{K}}$ as well. From census data at the ED level, we were able to estimate $\delta_{\bar{\bar{P}}=900}$, where 900 was the average number of persons in an ED, by a procedure described below. We then used these two calculated $\delta_{\bar{\bar{P}}}$'s to fit the curve $\delta_{\bar{\bar{P}}} = a\bar{\bar{P}}^b$ (p. 307 of Hansen, Hurwitz & Madow [6]). Having estimates of $\delta_{\bar{\bar{P}}}$ for only two values of $\bar{\bar{P}}$ is, of course, not very satisfactory, but we could do no better because of time restrictions. The computed values are as follows:

$$\delta_{\bar{\bar{P}}=3.1} = .55$$

$$\delta_{\bar{\bar{P}}=900} = .06$$

$$a \doteq .85$$

$$b \doteq -.4$$

With considerable effort, we were able to estimate $\delta_{\bar{\bar{P}}}$, the intraclass correlation between persons within ED's, from census data at the ED level. We needed some special tabulations from the 1970 decennial census, but time and money constraints prohibited running the entire census file, so calculations were made initially only for Wisconsin. (Calculations were subsequently made for Georgia and generally confirmed the earlier results.) There probably are some significant differences between the intraclass correlations for some States and those for Wisconsin and there may also be nontrivial changes in the intraclass correlations from 1970 to 1976, though these latter differences would not necessarily affect the optimum noncompact cluster size.

From the 20-percent census[4] data at the ED level, we computed:

$$\delta_{\bar{\bar{P}}=900} = \frac{S_1^2 - \bar{\bar{K}}\,S_2^2}{S_1^2 + \bar{\bar{K}}(\bar{\bar{K}}-1)\,S_2^2} \qquad (4)$$

where $S_1^2 = \frac{1}{M_s-1}\left\{\sum_{i=1}^{M_s} X_i^2 - \frac{1}{M_s}\left(\sum_{i=1}^{M_s} X_i\right)^2\right\}$

$$S_2^2 = \frac{1}{K_s}\sum_{i=1}^{M_s}\frac{X_i(K_i-X_i)}{K_i-1}$$

K_s is the State population,

M_s is the number of ED's in the State,

$\bar{\bar{K}} = \frac{K_s}{M_s}$, the average number of persons per ED,

K_i is the population of the i^{th} ED,

and X_i is the number of poverty children in the i^{th} ED.

The final quantities needed for computing the design effect for compact clusters (Formula (1)) are $(\hat{V}_K^2/V_K^2)$ and $(\hat{V}_L^2/V_L^2)$. Both these ratios are are assumed as constants in the calculations. In fact, however, they are somewhat a function of the cluster size. For $(\hat{V}_L^2/V_L^2)$ fewer clusters mean a larger number of ED's would turn out to be self-representing and would not contribute to (V_L^2/V_L^2). (In the extreme where all ED's are self-representing, $(\hat{V}_L^2/V_L^2)=1.0$, otherwise it is greater than 1. Since ED's are selected with a probability based on their size, (V_L^2/V_L^2) is expected to be close to 1.0.) For $(\hat{V}_K^2/V_K^2)$, the relative variation in number of persons per cluster is likely to decrease with the size of the cluster since large households will be combined with small households in larger clusters. Also, for characteristics of a small proportion of the total population, this quantity is not appreciably affected by the size of the cluster. (V_L^2/V_L^2) is assumed as a constant and thus left out of the

computations entirely. $(\hat{V}_K^2/V_K^2)$ could have been treated similarly, but instead was speculated as a constant 1.3 and carried through in the computations. Design effects for different compact cluster sizes are given in table 2.

TABLE 2. DESIGN EFFECTS FOR DIFFERENT CLUSTER SIZES

Cluster Size and Type	Design Effect
1 Housing Unit	2.8
COMPACT:	
2 Housing Units	4.0
3 Housing Units	5.0
NONCOMPACT:	
2 Housing Units	3.0
3 Housing Units	3.2
4 Housing Units	3.5
5 Housing Units	3.7
6 Housing Units	3.9

For noncompact clusters, $\delta_{\bar{N}}$ was estimated by[5]

$$\delta_{\bar{N}} = \frac{[1+\delta_{\bar{P}=900}(\bar{K}-1)] - [1+\delta_{\bar{P}=3.1}(\bar{K}-1)]}{(\bar{N}-1)\ [1+\delta_{\bar{P}=3.1}(\bar{K}-1)]} \qquad (5)$$

where

$\bar{N}$ is the average number of housing units per ED. Other terms were defined earlier.

The calculations yield $\delta_{\bar{N}} = .08$. Using this value in Formula (2) resulted in the design effects for noncompact clusters also shown in table 2.

We decided that any cost advantages for compact versus noncompact clusters were not sufficient to make up for the sizable design effect differences as shown in table 2, and then proceeded to determine the optimal cluster size for noncompact clusters. Table 3 compares the variable costs for additional interviews and interviewers to be incurred for alternative cluster sizes for a particular portion of the country. (Total survey costs run much higher.) The portion represented is the full States of Maryland and Massachusetts, and Milwaukee, Dane, and Brown counties in Wisconsin. The main reason for the choice of these particular areas is that data for direct field costs happened to be readily available for them.

Calculations were made for each cluster size in each of the five areas separately, and then summed to produce table 3. Consider Maryland, for example. For a given cluster size, the appropriate design effect was used to determine the number of sample units needed to achieve a CV of 10 percent on the estimated number of children in poverty families. The 1970 census figure on children in poverty families was used for the level of the estimate. The number of interviewers required for such a sample size was then estimated, which in turn determined the cost of training and recruiting. It was assumed, based on prior survey experience, that an additional interviewer is required for each increase of 100 units in sample. The training and recruiting cost was $350 per interviewer. Sampling costs were mostly a function of the number of ED's in sample. Field costs represent the direct interviewing costs. The table shows that the cost was minimized for clusters of three and thus this is what we used in the actual survey.

For the three counties in Wisconsin, we determined the number of sample units required for the State as a whole and then allocated this down for the three counties of interests. Cost figures were then developed separately for each county in the same manner as for the two States.

We made another set of computations that tended to confirm the estimated $\delta_{\bar{P}=3.1}=.55$ and $(V_K^2/V_K^2)=1.3$. These computations were also based on the special tabulation from the 1970 census giving the distribution of families with 0 children in poverty, 1 child in poverty, etc. From this distribution, we calculated an estimate of the population relvariance between households of the number of children aged 5-17 in poverty per household; this relvariance is V_L^2 defined previously. We also determined the population relvariance between persons for the proportion of total persons that are poverty children aged 5-17. This is the same as V_K^2 defined previously where $V_K^2 = (1-P)/P$, where P is the proportion of poverty children.

To compare V_L^2 with V_K^2, which can be considered as the relvariances for a simple random sample of one household and one person respectively, V_L^2 needs to be adjusted by the average number of persons per household. Therefore the design effect for a simple random sample of n_H households versus a simple random sample of $\bar{K}n_H$ persons, with $\bar{K}=3.1$, is given by, $\bar{K}\ V_L^2/V_K^2$. This design effect was approximately 2.8. Although this procedure involves less computations; it does not produce a value for the intraclass correlation between persons within a household that was used in other aspects of the sample design.

IV. DURBIN-SAMPFORD SAMPLE SELECTION

As stated previously, the prime objective of SIE was to produce estimates of children, age 5-17, in poverty families with coefficients of variation no worse than 10 percent for each State. It was felt that reliable estimates of variance were desirable in order to verify that we had met the specified reliability requirements and that good estimates of variance be available for the analysis of the data resulting from the survey. The sample selection can be divided into two stages: First, the selection of PSU's and second the selection of housing units from PSU's. The discussion will be divided between these two stages.

Selection of Primary Sampling Units. Generally when there is sufficient auxiliary information (usually from the most recent census) to enable us to stratify the PSU's, it is assumed that only one PSU needs to be selected from each stratum. Under this assumption of sufficient auxiliary information, the between PSU component of variance is felt to be smaller when forming small strata from which one PSU is selected than when forming larger strata from which more than one PSU is selected. Also, the method of estimating variance when one PSU per stratum has been selected, collapsed strata, produces biased estimates of variance. This bias may be large enough so that the expected value of the variance estimate for collapsed strata is greater than the unbiased estimate of variance for a design

specifying two PSU's per stratum and where the strata are larger than the one PSU per stratum design.

Two considerations entered into our decision whether to select one or two PSU's per stratum. First, the between PSU component of variance needed to be small enough so that the coefficient of variation for estimating poverty children remained less than 10 percent. Preliminary estimates indicated that the between PSU variance for two PSU's per stratum would range between 0 and 30 percent of the total variance (with three-quarters of the States below 10 percent) but that the coefficient of variation would remain below 10 percent for all States but one. Hence, using the Durbin [5] technique to select two PSU's per stratum would not raise the variances above the stated requirements. Second, we found that since the between PSU variance was a minor component of the total variance, the risk was small that we would estimate a coefficient of variation greater than 10 percent when it actually was less. If the risk of estimating a coefficient of variation greater than 10 percent had been high, we might have selected the procedure with the lowest variance since one PSU per stratum or two PSU's per stratum would have both been relatively unattractive. From this preliminary analysis, we concluded that there was no strong reason to prefer one PSU per stratum over two PSU's per stratum, and, as a result, we selected the procedure which would provide unbiased estimates of variance.

Table 4 illustrates an interesting relationship between the between PSU variance for Durbin technique, for one PSU per stratum, and for the collapsed stratum estimate of variance. These calculations were performed subsequent to the decision to use Durbin technique to select PSU's and had no bearing on that decision. Table 4 shows estimates from census data of the between PSU variance based on stratification by 1970 poverty children for six States. Part a. shows the between PSU variance for estimates of 1970 poverty children, that is, an item with perfect correlation with the auxiliary information. The expected relationship is evident in this part of the table, that one PSU per stratum is superior to two PSU's per stratum in terms of true variance but that the expected value of the collapsed stratum estimate of variance exceeds the true variance for two PSU's per stratum. Parts b. and c. of the table show the between PSU variance for 1970 poverty families, an item highly correlated with the auxiliary information, and for 1960 poverty families, an item that shows the effect of displacement in time. (Note that the number of 1960 poverty children is not available for comparison.) Part c. no longer shows a clear advantage for one PSU per stratum over two PSU's per stratum from larger strata. The collapsed stratum estimate of variance still provides an overestimate of the true variance though its relative overestimate is less. Taking into consideration the fact that the auxiliary information will not be perfectly correlated with the key items of a survey, Table 4 indicates that two PSU's per stratum may be a good compromise between the reduction in variance due to stratification and obtaining an unbiased estimate of variance.

Selection Within PSU's. The most common method used at the Bureau of the Census to select a sample of units from within a PSU is to take a sorted systematic sample. This method has the obvious advantages of being easy to implement and of resulting in a relatively low true variance for items correlated with the sort variables, but has the disadvantage that, since the systematic sample is in effect a sample of one cluster per PSU, no unbiased estimate of the variance exists. The methods to estimate the variance will tend to overestimate it when the sort variables are effective in reducing the variance.

Along with an estimate of variance due to the selection of PSU's as described above, we required an estimate of the within sampling variance for the self-representing sample PSU's and, since we were using the Durbin technique to select the non-self-representing PSU's, we required an estimate of the within sampling variance for each of the non-self-representing PSU's. We felt that the application of the Durbin technique to the within PSU sample selection would lead to little, if any, increase in variance over a systematic sample[6] since (1) the strata would be formed in the same sort as if a systematic sample were to be used, and (2) the strata would be numerous and small so that each stratum would be fairly homogeneous. Also, selecting a sample or estimating variances using the Durbin technique is not much more difficult on the computer than other procedures. Thus, we decided to select our sample from the 1970 census using the Durbin technique so that we would be able to estimate the gain in variance due to the sorting and stratification of the sample.

Approximately 85 to 90 percent of our sample in a State came from the 1970 census file and the Durbin technique was used to select the sample. The remainder of the universe, primarily new construction, required clerical operations for the sample selection. As a result, we chose systematic samples from this part of the universe. Since a systematic sample would provide little gain in variance for this part of the sample, the variance was estimated as if it were based on a simple random sample.

The sample selection within a PSU was briefly sketched in the introduction. As described in Section III, above, we decided to select a sample of ED's from which a noncompact cluster of three housing units would be selected from each one. Many ED's were large enough so that we expected them to enter sample with certainty. As a result, we decided to directly select a sample of housing units from these large ED's using the Durbin technique. The housing units from large ED's were sorted by their poverty level and the number of children under 18, and within these by county and ED. In this sort, the housing units were grouped into strata (called Durbin housing unit strata), and two units per stratum were selected using the Durbin-Sampford rejective method.

The remaining smaller ED's were sorted by five size categories and, within each size category, by the percent of persons in poverty such that the first size category was sorted from highest to lowest poverty, the second size category from lowest to highest poverty, etc. In this sort, the ED's were grouped into 12 or more strata

(called Durbin ED strata), and two ED's per stratum were selected using the Durbin-Sampford rejective method where the measure of size was proportional to the number of housing units plus the number of persons in special places divided by three. From a selected ED either a special place or a cluster of three housing units was selected. In either case a systematic sample was taken. It was thought that a substantial improvement in variance could be achieved if the housing units within an ED were sorted by poverty level and number of children less than 18, and a systematic sample of three housing units was taken.

Departures from an Unbiased Estimate of Variance. Approximate methods need to be applied to estimate the variance from those frames from which systematic samples were selected. Thus, though we attempted to select the sample so that we would have an unbiased estimate of variance, we did not achieve this fully. Furthermore, we departed from an unbiased estimate of variance in two additional ways.

First, an estimate of variance for a non-self-representing stratum k is:

$C_k(\hat{X}_{k1}-\hat{X}_{k2})^2 + (1-C_k)\ (\hat{\sigma}^2_{k1} + \hat{\sigma}^2_{k2})$ where

$\hat{X}_{k1}$ and $\hat{X}_{k2}$ are sample estimates from the sample PSU's from stratum k,

$\hat{\sigma}^2_{k1}$ and $\hat{\sigma}^2_{k2}$ are estimates of the within sampling variance for the two PSU's, and

C_k is a constant which is a function of the joint probability of selecting the two sample PSU's and of the probabilities of selecting each of the PSU's in stratum k.

Durbin [5] recommends that the coefficient C_k in the variance formula be reduced to one whenever it exceeds one in order to reduce the variance on the variance estimate. C_k exceeds one when the measures of size of the units in a stratum are diverse and one of the larger units is not in the pair of selected units. Since, the measures of sizes of PSU's in a stratum were rather heterogeneous for this survey, the C_k's were reduced to one according to Durbin's recommendation.

Table 5 shows estimates of the relative bias due to reducing C_k to one, the relative variance of the unbiased estimate of variance, and the relative mean square error of the biased estimate of variance for estimating 1970 poverty children for six States. The relative bias is in general small and there is a decrease in the relative mean square error.

Second, the selection of census housing units was from the 1970 census 20-percent sample. By using the Durbin method to select the sample of housing units from large ED's in the first method of selection, an estimate of the within component of variance due to the 20-percent census sample was required. For the State of Wyoming with an overall sampling rate of about 1 in 25, ignoring this within component of variance would produce a substantial underestimate of the variance (approximately 16 percent). Thus, we decided to estimate variances under the assumption that we had selected a stratified simple random sample from all units represented by the 20-percent census sample. This procedure can be defended as follows: The Durbin sampling from the 20-percent census sample is approximately simple random sampling because the measures of size were, in general, nearly equal. If we assume that the 20-percent census sample was a simple random sample, then we can conclude that, overall, we had selected a simple random sample of households.

Reduction in Sample. Section II of this paper discusses how the sample was reallocated due to the budget-imposed reduction. There was no reduction in nine States and a reduction from approximately 2 to 24 percent in the remaining States. Because we had selected our sample of ED's or of housing units using the Durbin method, the reduction in sample was complicated.

The reduction in cost could be achieved in two ways: (1) by eliminating interviews and (2) by reducing the number of ED's an enumerator would have to visit. Thus we could reduce the cost both ways if we deleted every sample housing unit from an ED. Reducing ED's from the Durbin ED selection was no problem. A systematic sample of Durbin ED strata was selected and one of the two sample ED's was randomly deleted with equal probability. It was thought that the increase in variance due to deleting one of two ED's from a stratum was less than the increase in variance from deleting both ED's from half as many strata; no estimates were made of this difference.

For the housing unit selection using the Durbin procedure, it was thought that there could be an excessive increase in variance if all housing units in a sample ED were deleted because of the clustering of the sample housing units. Table 6 shows for States in which a large part of the housing units had been directly selected using the Durbin procedure, the average number of sample households per ED, the approximate percent reduction required, the increase in variance of an ED reduction over a simple housing unit reduction, and the expected CV's after an ED reduction. Four States had an excessive increase in variance from an ED reduction that brought their expected CV's over 10 percent. For each of these four States (Delaware, Wyoming, Nevada, and Hawaii), a systematic sample of Durbin housing unit strata was selected and both housing units from a selected stratum were deleted for the reduction. For the remaining States, the ED's with sample housing units selected using the Durbin procedure were ordered by the number of sample housing units in the ED and a systematic sample of ED's was deleted for the reduction. Note that, because of the underlying random structure of the Durbin sample selection, every pair of ED's from the Durbin ED selection and every pair of housing units from the Durbin housing unit selection retains a positive joint probability of selection in spite of the systematic reduction. Hence, an unbiased estimate of variance (except as noted above) can still be obtained. The derivation of the unbiased estimates of variance has been completed but their presentation would be rather complicated and they will not be included in this paper. Documentation has not been completed.

Advantages and Disadvantages of the Durbin-Sampford Selection Method. The Durbin-Sampford method of sampling selection has been discussed in two contexts in this paper. First, the selection of PSU's, comparing the Durbin procedure with one PSU

per stratum, and second the selection from within sample PSU's, comparing the Durbin procedure with systematic sampling.

The disadvantages of using Durbin procedure, with respect to one PSU per stratum and systematic sampling, are approximately the same. First, it is more difficult to select the sample using the Durbin-Sampford method. This is rather small when it is implemented on the computer but it could be a very difficult task when selecting the sample by hand if there were to be numerous strata. Second, the estimate of variance is more complicated since a constant for each stratum has to be calculated and additional components of variance usually need to be estimated. This increase in the difficulty of estimating variances can often be reduced if the variance estimate can be adapted to replications. Durbin [5] points out a method to estimate the variance which can be easily adapted to the replication method of estimating variances. Third, the true variance from the Durbin procedure may be larger when an item is highly correlated with the auxiliary information that was used for sorting or stratification. Finally, a sample selected using the Durbin procedure is less versatile if a further supplementation or reduction in the sample is required.

The most obvious advantage of the Durbin method is that it provides an unbiased estimate of variance. It appears to be a reasonable balance between providing an unbiased estimate of variance and reducing variance by sorting and stratification of the universe. This would in general be true for any scheme which selects two units per stratum. The advantage of the Durbin procedure over other without replacement schemes is that the Durbin-Sampford rejective method is comparatively easy to implement and that the constants in the variance estimate are directly calculable from terms used in the sample selection procedure. Sampling with replacement is easier to implement but will produce variances larger than Durbin.

A second less obvious advantage for the Durbin method is that the estimate of variance may itself have a lower variance than the estimated variance from the collapsed stratum technique. This can be argued as follows. An estimate of total variance from Durbin selection is:

$$\sum_{k}^{NSR} C_k(\hat{X}_{k1}-\hat{X}_{k2})^2 + \sum_{k}^{NSR} (1-C_k)(\hat{\sigma}^2_{k1}+\hat{\sigma}^2_{k2}) + \hat{\sigma}^2_{SR}, \qquad (1)$$

($\hat{\sigma}^2_{SR}$ is an estimate of the within sampling variance for the self-representing PSU's). The second and third terms, for our survey, have a considerable lower variance than the first terms since they are made up of the sum of squares from numerous strata and the first term is made up of at most 10 sum of squares. Now the expected value of the first term is:

$$\sum_{k}^{NSR} E\, C_k(\hat{X}_{k1}-\hat{X}_{k2})^2 = \sum_{k}^{NSR}\sum_{i>j}\sum(\pi_{ki}\,\pi_{kj}-\pi_{kij})(X_{ki}-X_{kj})^2 + \sum_{k}^{NSR}\sum_{i}\pi^2_{ki}\sigma^2_{ki} \qquad (2)$$

where π_{ki} and π_{kj} are the probabilities of selecting PSU_i and PSU_j from stratum k, π_{kij} is the joint probability of selecting both PSU's i and j, and $X_{ki}=E(\hat{X}_{ki}|\text{the selection of } PSU_i)$. Thus, the first and least accurate term estimates the between PSU component of variance and part of the within component of variance for the non-self-representing stratum. The remainder of the within variance is estimated by the more accurate second and third terms. Similarly, the estimate of total variance from collapsed stratum is $\sum_{k}^{\text{Paired Strata}} (\hat{X}_{k1}-\hat{X}_{k2})^2 + \hat{\sigma}^2_{SR}$ and the expected value of the first term

$$\sum_{k}^{\text{Paired Strata}} E(\hat{X}_{k1}-\hat{X}_{k2})^2 = \sum_{k}^{\text{Paired Strata 2}}\sum_{j=1}^{2}\sum_{i} P_{kji}(X_{kji}-\bar{X}_{kj})^2 + \sum_{k}^{\text{Paired Strata 2}}\sum_{j=1}^{2}\sum_{i} P_{kji}\sigma^2_{kji} + \sum_{k}^{\text{Paired Strata}} (\bar{X}_{k1}-\bar{X}_{k2})^2 \qquad (3)$$

where P_{kji} is the probability of selecting PSU_i from j^{th} strata of the paired stratum denoted by k.

$X_{kji} = E(\hat{X}_{kji}|\text{the selection of } PSU_i)$, and

$\bar{X}_{kj} = \sum_{i} P_{kji}X_{kji}$.

Thus the collapsed stratum term estimates the between PSU component of variance, the within component of variance for NSR strata and the bias from using the collapsed stratum estimator. Again the first term is the least accurate term in the estimate of variance since it would be made up of, at most, 10 sums of squares for SIE. Thus, the least accurate first terms in the estimates of variance, estimate more of the total variance for collapsed stratum than for the Durbin procedure. Thus, the variance estimate may be less accurate for collapsed strata. Research into the variance of our variance estimates will have to be conducted before a conclusive statement can be made that the Durbin procedure results in a lower variance on the variance estimate.

A final advantage is that the underlying Durbin structure allows us to estimate unbiasedly for this survey the increase in variance due to the sample reduction and the variance before the reduction. Thus, we will be able to evaluate the effect of the reduction on our estimates. The variances are currently being estimated.

FOOTNOTES

[1]The Current Population Survey is a monthly survey conducted by the Bureau of the Census for the Bureau of Labor Statistics. Its prime purpose is to produce monthly labor force data, but in March of each year an extensive set of supplementary questions on income and household composition are asked which makes possible estimates of children in poverty families. For more details, see Thompson [9].

[2]The basis for the formulae and methodology used is given in chapter 6 of Hansen, Hurwitz, and Madow [6]. Notation here is not generally consistent with the book.

[3]Strictly speaking, this formula was not used and the resultant data in table 2 was not produced in our earlier work. This is equivalent

to the earlier work, though, and is presented in this form for ease of comparison.

[4]Recall that the final sample was selected from the 20-percent census data.

[5]Page 267, Hansen, et al. [6]

[6]Cochran [3] has shown that if the population is autocorrelated, that is, $\rho_i \geq \rho_{i+1} \geq 0$, and the correlogram is concave upwards, that is $\rho_{i-1}+\rho_{i+1}-2\rho_i \geq 0$, then systematic sampling is superior to stratified sampling taking one unit per stratum, where ρ_i is the correlation between two units which are i units apart in a listing of the population. Because of the sort, described above, that was imposed on the census frame prior to the sample selection, the census population is autocorrelated for estimates of poverty, but no result has been derived that shows the conditions under which a systematic sample is superior to a stratified, two units per stratum without replacement.

REFERENCES

[1] Boisen, Morton. "Study Plan for Sample Design on Survey of Income and Education, Public Law 93-380." Internal Census Bureau memorandum to Daniel B. Levine, May 12, 1975.

[2] Boisen, Morton. "Sample Sizes and Sample Design Decisions for the Survey of Income and Education." Census Bureau memorandum to Wray Smith, Department of Health, Education and Welfare, July 18, 1975.

[3] Cochran, W. G. "Relative Accuracy of Systematic and Stratified Random Samples of a Certain Class of Populations." Annals of Mathematical Statistics, 17, (1946) pp. 164-177.

[4] Dippo, Cathryn. "Expansion of CPS to Provide Reliable State Estimates of Unemployment." Proceedings of the Social Statistics Section, American Statistical Association, 1975, pp. 387-391.

[5] Durbin, J. "Design of Multi-Stage Surveys for the Estimation of Sampling Errors." Applied Statistics, 16, (1967) pp. 52-46.

[6] Hansen, Morris H., Hurwitz, William N., and Madow, William G. Sample Survey Methods and Theory, Vol. 1. New York: John Wiley & Sons, 1953.

[7] Sampford, M. R. "On Sampling Without Replacement with Unequal Probabilities of Selection." Biometrika, 54, (1967) pp. 499-513.

[8] Smith, Wray. "Reduction of the Supplemental Sample for the SIE." Department of Health, Education, and Welfare memorandum to Earle J. Gerson, Census Bureau, Dec. 7, 1975.

[9] Thompson, Marvin M. and Shapiro, Gary. "The Current Population Survey: An Overview." Annals of Economic and Social Measurement, Vol. 2, No. 2, (April 1973) pp. 105-129.

[10] U. S. Congress, Elementary and Secondary Education Act of 1965, Pub. L. 89-10, 89th Cong., 1st sess., 1965, H.R. 2362.

[11] U. S. Congress, Education Amendments of 1974, Pub. L. 93-380, 93rd Cong., 2nd sess., 1974, H.R. 69.

ACKNOWLEDGMENTS

The sample design for this survey resulted from the contribution of many people besides the authors. Of particular importance are Wray Smith, David Bateman, Paul Bettin, Harold Nisselson, and William P. Smith, III. Mr. Bettin also contributed to the writing of Section III of this paper. Helpful comments were made by Wray Smith, Paul Bettin, George H. Gray, Charles D. Jones, Henry Woltman, and an anonymous Census Bureau referee. The typing was done by Edith Oechsler, assisted by Arlene Sagin.

TABLE 3. COSTS AND SAMPLE SIZES FOR ALTERNATIVE NONCOMPACT CLUSTER SIZES

Noncompact Cluster Size	No. of Sample Units Required for 10% CV	No. of Interviews Required Above Minimum (For Cluster Size of 1)	Cost of Training and Recruiting Additional Interviewers	No. of Sample ED's	Variable Cost of Sample Selection	Direct Field Costs	Sum of (6) and (7)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	9,673	--	--	9,673	$34,000	$41,400	$75,400
2	10,343	6	$ 2,100	5,171	18,000	41,305	61,405
3	11,110	14	4,900	3,703	14,000	40,815	59,715
4	11,876	22	7,700	2,969	13,200	41,738	62,638
5	12,644	29	10,150	2,529	13,000	43,092	66,242
6	13,410	37	12,950	2,236	12,000	45,186	70,136

TABLE 4

Between PSU Variance for Selected States for Estimates Based on Stratification by 1970 Poverty Children

	State and Characteristic	2 PSU's per stratum (Durbin)	1 PSU per stratum	Collapse Strata		Number of sample NSR PSU's
				Unadjusted	Adjusted[1]	
		(000)	(000)	(000)	(000)	
a.	Estimated 1970 Poverty Children	(1)	(2)	(3)	(4)	
	Alabama	39,984	18,843	65,671	66,286	12
	California	22,839	10,181	101,697	62,260	8
	Florida	36,182	14,448	64,168	62,866	8
	Michigan	2,380	853	4,167	4,829	12
	South Dakota	2,880	767	4,264	4,245	14
	Washington	918	188	3,940	2,648	8
b.	Estimated 1970 Poverty Families					
	Alabama	11,198	12,639	22,417	27,653	
	California	8,405	7,268	25,835	16,105	
	Florida	23,862	16,075	38,783	37,414	
	Michigan	2,858	2,908	3,174	3,603	
	South Dakota	544	504	818	968	
	Washington	639	448	1,810	1,183	
c.	Estimated 1960 Poverty Families					
	Alabama	44,523	52,535	78,961	96,947	
	California	20,126	22,022	49,029	34,606	
	Florida	61,388	34,726	121,144	107,063	
	Michigan	15,408	15,615	17,844	19,735	
	South Dakota	2,987	2,804	4,576	5,381	
	Washington	1,865	1,922	4,032	3,020	

[1]The adjusted collapsed strata estimate attempts to reduce the bias by eliminating the bias due to the different strata size. For this column, the following estimate was used:

$$(a_1\hat{X}_1 - a_2\hat{X}_2)^2$$

where $a_1 = \frac{2N_2}{N_1+N_2}$ and $a_2 = \frac{2N_1}{N_1+N_2}$,

N_1 and N_2 are the 1970 populations in the two paired strata.

TABLE 5

Bias and Relative Mean Square Error for Selected States in the Adjusted Durbin Estimate of Variance for 1970 Children in Poverty

STATE	Variance Due to the Selection of Primary Sampling Units	Bias in Reducing C_K to 1	Total Variance	Relative Bias Due to Reducing C_K to 1	Relative Variance of the Unbiased Variance Estimate	Relative MSE of the Biased Variance Estimate
	(1) (000)	(2) (000)	(3) (000)	(4) (000)	(5) (000)	(6) (000)
Alabama	39,984	-247	393,801	-0.0006	0.16254	0.15457
Georgia	55,623	-176	638,425	-0.0003	0.19655	0.19189
North Dakota	272	- 90	3,999	-0.0225	0.16848	0.06416
Ohio	4,849	- 5	540,310	-0.0000	0.01798	0.01731
South Dakota	2,880	- 21	12,894	-0.0017	0.23600	0.22448
Utah	945	- 34	6,402	-0.0053	0.07791	0.06808

TABLE 6

Increase in Variance of an Estimate of 1970 Children in Poverty Due to Deleting ED's from the Durbin Housing Unit Sample

STATE	Average Number of Sample HU's Per ED	Approximate Percent Reduction Required	Increase in Variance Over Housing Unit Reduction[1]	Expected Coefficient of Variation from an ED Reduction (%)
	(1)	(2)	(3)	(4)
Maine	2.8	5	1.009	9.4
New Hampshire	NA	2	1.023	9.9
Vermont	5.2	4	1.056	10.1
Rhode Island	NA	5	1.024	9.9
Connecticut	3.1	8	1.006	9.8
Nebraska	3.0	6	1.030	9.8
Kansas	2.9	6	1.025	9.8
Delaware	6.4	20	1.206	10.8
District of Columbia	2.9	10	1.023	9.7
Montana	4.3	7	1.057	9.9
Idaho	NA	5	1.085	10.2
Wyoming	7.7	14	1.770	13.0
New Mexico	3.7	18	1.041	7.0
Nevada	5.7	8	1.136	10.4
Hawaii	5.9	24	1.144	10.5

[1]These were derived from a regression model.

VOTING RIGHTS SURVEY

Grant Capps, U.S. Bureau of the Census

As is well known, the design of any survey will generally involve certain assumptions and "guesstimates" regarding various unknown parameters (frequently costs and variances). The accuracy and amounts of these assumptions will usually depend upon the available funds, lead-time, and prior information. As this paper demonstrates for the 1976 Voting Rights Survey, designed and conducted by the U.S. Bureau of the Census, there were sufficient resources available so as to reduce the usual educated guesstimating associated with efficient survey design.

I. INTRODUCTION

A. Survey Background. The 1976 Voting Rights Survey was concerned with measuring the voting participation rates for certain minorities in specified jurisdictions scattered across the nation. Congress, the Department of Justice, and the Census Bureau jointly identified 93 jurisdictions to be surveyed. These jurisdictions consisted of 11 towns, 73 counties, and 9 States. The minorities, which varied by jurisdiction, included the Black, Spanish, American Indian, Japanese, Chinese, Filipino, and Native Alaskan ethnic groups. Depending upon the costs involved, either a complete census or sample survey was conducted within each jurisdiction. Enumeration occurred within 6 months of the November 1976 presidential election. The results of the survey are expected to be available by November 1977.

B. Purpose of Paper. The purpose of this paper is to describe the major aspects and considerations involved in the sample design for the 1976 Voting Rights Survey. In doing so, the necessary theory will be developed along with the assumptions involved, and the relevant results will be given. The following major survey design problems and their solutions will be presented at length:

1. The determination of the increase in the variance of the estimated minority voting rate due to the clustering of people within households.
2. The determination of both (a) the increase in the variance due to the clustering of housing units, and (b) the optimum cluster size.
3. For each statewide jurisdiction, the determination of a variance function explicitly denoting the components of variance due to (a) the selection of primary sampling units (PSU's), usually counties, and (b) the subsampling within the chosen PSU's.
4. For each statewide jurisdiction, the joint determination of the optimum combination of within PSU sample size, number of sample PSU's, PSU measure of size, and (within PSU) cluster size.

Certain other relatively straightforward aspects of the sample design, such as the allocation of the sample to the various strata, will also be discussed, but to a lesser extent.

C. Survey Requirements. The survey was designed so that the estimated minority voting rate within each jurisdiction would have about a 10% coefficient of variation (CV). For each identified jurisdiction, all minorities which comprised 5% or more of the 18+ population in the jurisdiction were, by definition, minorities of interest. In those jurisdictions with more than one minority of interest, the 10% CV reliability requirement was applied separately to each such minority. In 30 jurisdictions (all the towns and 19 counties), the estimated cost for a complete census was less than that of a comparable sample survey designed to meet the 10% CV requirement. Thus, in these 30 census jurisdictions, the estimates will be free of sampling error, although nonsampling error will be present.

II. INTRACLASS CORRELATIONS AND DESIGN EFFECTS FOR WITHIN COUNTY SAMPLING

In order to determine the approximate sample size needed to meet the 10% CV reliability requirement in the designated counties, it was first necessary to estimate the variance effects of clustering for both persons and housing units. This section reviews the relevant theory and describes the method with which it was applied in determining an appropriate variance model for sampling within the county jurisdictions. As will be shown later, the conclusions arrived at in this section will also be employed when developing the variance formulae pertaining to the statewide jurisdictions.

A. Notation and Definitions. Consider the following situation in a typical county jurisdiction. Let there be M clusters (primary units) of housing units (listing or secondary units), with the i^{th} cluster containing N_i housing units (HU's) for a total of $N = \sum_{i}^{M} N_i$ HU's. The j^{th} HU in the i^{th} cluster contains K_{ij} people 18 and over (elementary units) for a total of

$K_i = \sum_{j}^{N_i} K_{ij}$ 18+ persons in the i^{th} cluster and

$K = \sum_{i}^{M} K_i$ 18+ persons in the entire county.

Let $\bar{N} = N/M$ and $\bar{\bar{K}} = K/N$. The sampling plan we wish to consider involves the selection of m sample clusters (primary units) followed by the secondary selection of $n_i (\leq N_i)$ sample HU's in the i^{th} selected cluster. Let k_{ij} denote the number of 18+ sample persons in the j^{th} sample HU of the i^{th} sample cluster. Assume that simple random sampling is used at both stages and further, that the second stage sampling fraction $f_2 = n_i/N_i$ is constant for all i. The expected total sample size is $n = E\left[\sum_{i}^{m} n_i\right] = E\left[\sum_{i}^{m} f_2 N_i\right] = m\bar{N}f_2$ HU's and the average number of sample HU's per sample cluster is $\bar{n} = \frac{n}{m} = \bar{N}f_2$ HU's. All persons within a sample HU will of course be interviewed and thus $k_{ij} = K_{ij}$. The expected sample of 18+ sample people is $k = E\left[\sum_{i}^{m}\sum_{j}^{n_i} k_{ij}\right] = E\left[\sum_{i}^{m}\sum_{j}^{n_i} K_{ij}\right] = n\bar{\bar{K}}$ and the average number of 18+ sample people per sample HU is $\bar{\bar{k}} = k/n = \bar{\bar{K}}$.

Let:

$$Y_{ij\ell} = \begin{cases} 1 \text{ if the } \ell^{th} \text{ person in the } j^{th} \text{ HU of cluster } i \text{ is an 18+ minority of interest citizen.} \\ 0 \text{ if not.} \end{cases}$$

and

$$X_{ij\ell} = \begin{cases} 1 \text{ if } Y_{ij\ell} = 1 \text{ and the } ij\ell^{th} \text{ person voted.} \\ 0 \text{ if not.} \end{cases}$$

Population Totals: $Y_{ij} = \sum_{\ell}^{K_{ij}} Y_{ij\ell}$, $Y_i = \sum_{j}^{N_i} Y_{ij}$, $Y = \sum^{M} Y_i$,

Population Means: $\bar{\bar{Y}}_{ij} = \frac{Y_{ij}}{K_{ij}}$, $\bar{\bar{Y}}_i = \frac{Y_i}{N_i}$, $\bar{Y} = \frac{Y}{M}$, $\bar{\bar{Y}} = \frac{Y}{N}$, $\bar{\bar{Y}} = \frac{Y}{K}$.

Similarly define the corresponding population totals and means for the variate $X_{ij\ell}$. The unknown parameter to be estimated by the sample is the minority voting rate $R = X/Y$.

Similar definitions can be attached to the sample quantities by simply replacing the upper case letters with lower case ones. For example, we have:

$$y_{ij\ell} = \begin{cases} 1 \text{ if the } ij\ell^{th} \text{ sample person is an 18+ minority of interest citizen.} \\ 0 \text{ if not.} \end{cases}$$

$$x_{ij\ell} = \begin{cases} 1 \text{ if } y_{ij\ell} = 1 \text{ and the } ij\ell^{th} \text{ sample person voted.} \\ 0 \text{ if not.} \end{cases}$$

and

$y_{ij} = \sum_{\ell}^{k_{ij}} y_{ij\ell}$, $y_i = \sum_{j}^{n_i} y_{ij}$, and $y = \sum^{m} y_i$.

Unbiased estimators for Y and X are $y' = \frac{N}{n} y$ and $x' = \frac{N}{n} x$, respectively. Thus, to estimate the minority voting rate R, we use the (nearly) unbiased estimator $r = \frac{x'}{y'} = \frac{x}{y}$. And, finally, let us define one more set of variables. Let $U_{ij\ell} = X_{ij\ell} - R Y_{ij\ell}$. Define the population totals and means corresponding to $U_{ij\ell}$ exactly as with $Y_{ij\ell}$ and $X_{ij\ell}$.

That is, $U_{ij} = \sum_{\ell}^{K_{ij}} U_{ij\ell}$, $U_i = \sum_{j}^{N_i} U_{ij}$, $U = \sum^{M} U_i = 0$ and similarly for the various population means (note $\bar{U} = \bar{\bar{U}} = U = 0$).

For the sample quantities we again use lower case letters. Begin with $u_{ij\ell} = x_{ij\ell} - R\, y_{ij\ell}$. Note that unlike $y_{ij\ell}$ and $x_{ij\ell}$, $u_{ij\ell}$ is an unobservable random variable. The sample totals are

$u_{ij} = \sum_{\ell}^{k_{ij}} u_{ij\ell}$, $u_i = \sum_{j}^{n_i} u_{ij}$, and $u = \sum^{m} u_i = x - Ry$, with obvious definitions for the sample means. An unbiased estimator for U=0 (but certainly not a statistic) is, of course $u' = x' - Ry'$. u' is involved in the Taylorized form for r.

B. Expressions for the Relative Variance of r.

For the above sampling scheme one can refer to nearly any sampling text and obtain the following straightforward approximation for the relative variance of r:

$$V_r^2 = \frac{\mathrm{Var}(r)}{R^2} \doteq \frac{\mathrm{Var}(x' - Ry')}{R^2 Y^2} = \frac{\mathrm{Var}(u')}{X^2} \tag{1}$$

$$= \frac{M-m}{M} \frac{B^2}{m} + \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{W^2}{m\bar{n}} \tag{2}$$

where,

$$B^2 = \frac{\sum^{M} (U_i - \bar{U})^2}{(M-1)\bar{\bar{X}}^2} \tag{3}$$

$$= \frac{\sum^{M} (X_i - RY_i)^2}{(M-1)\bar{\bar{X}}^2} \tag{4}$$

and

$$W^2 = \frac{1}{N\bar{\bar{X}}^2} \sum_{i}^{M} \frac{N_i}{N_i - 1} \sum_{j}^{N_i} (U_{ij} - \bar{\bar{U}}_i)^2 \tag{5}$$

$$= \frac{1}{N\bar{\bar{X}}^2} \sum_{i}^{M} \frac{N_i}{N_i - 1} \left[\sum_{j}^{N_i} (X_{ij} - RY_{ij})^2 - \frac{1}{N_i} (X_i - RY_i)^2 \right] \tag{6}$$

The first term in (2) is the familiar between-cluster component of relative variance and the second term is the within cluster component which obviously vanishes if there is no cluster subsampling, i.e., if $\bar{n} = \bar{N}$ or $f_2 = 1$.

Since it is desired to express V_r^2 in terms of known or easily guessimated parameters, it is necessary to modify (2). The best reference for accomplishing such a modification is chapter 6, volume 1, of the Hansen, Hurwitz & Madow [3] sampling text. It is stated there (p. 264) that (2) is very nearly equal to

$$V_r^2 = \left[\frac{1-f}{m\bar{n}} V_L^2 \right] \frac{\hat{V}_L^2}{V_L^2} [1 + \delta_L (\bar{n} - 1)], \tag{7}$$

where, $f = \frac{m}{M} f_2 = \frac{n}{N}$, and

$$V_L^2 = \frac{\sum_{i}^{M} \sum_{j}^{N_i} (X_{ij} - RY_{ij})^2}{(N-1)\bar{\bar{X}}^2} \tag{8}$$

$$\hat{V}_L^2 = \frac{M-1}{M} B^2 + \frac{\bar{N}-1}{\bar{N}} W^2 \tag{9}$$

$$= V_r^2 (m = \bar{n} = 1) \tag{10}$$

and

$$\delta_L \doteq \frac{B^2 - \frac{W^2}{\bar{N}}}{\hat{V}_L^2} \text{ (for large M).} \tag{11}$$

The subscript L denotes the listing unit, which here is the HU. The first term in brackets in (7) is the relative variance for a simple random sample of $m\bar{n}$ HU's. The unbracketed middle term of (7) is a factor which should be just slightly greater than unity and is present only if N_i varies from cluster to cluster. Finally, δ_L is the intraclass or intraclass correlation among HU's within clusters of HU's. δ_L is a measure of the homogeneity or similarity among HU's in the same cluster and satisfies $-\frac{1}{\bar{N}-1} \leq \delta_L \leq 1$, taking on the value one when there is perfect within primary unit homogeneity.

V_L^2 in the first term in brackets of (7) can be eliminated by simultaneously applying both (2) and (7) in the case of a single stage simple random sample of mn=n HU's and equating the results. Using form (2) for a random sample of n HU's yields:

$$\frac{N-n}{N}\,\frac{\sum_i^M \sum_j^{N_i} (X_{ij}-RY_{ij})^2}{n\,\bar{\bar{X}}^2(N-1)} = \frac{1-f}{n}\,V_L^2, \qquad (12)$$

and applying expression (7) gives

$$\left[\frac{1-\frac{n}{N}\frac{\bar{\bar{k}}}{\bar{\bar{K}}}}{n\bar{\bar{k}}}\,V_2^2\right]\frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{k}}-1)]$$

$$=\left[\frac{1-f}{n\bar{\bar{K}}}\,V_2^2\right]\frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{K}}-1)], \qquad (13)$$

where

$$V_2^2 = \frac{\sum_i^M \sum_j^{N_i} \sum_\ell^{K_{ij}} (U_{ij\ell}-\bar{\bar{U}})^2}{(K-1)\,\bar{\bar{X}}^2} \qquad (14)$$

$$\doteq \frac{1-R}{R\,\bar{\bar{Y}}} \qquad (15)$$

$$B_2^2=V_L^2,\quad W_2^2=\frac{1}{K\bar{\bar{X}}^2}\sum_i^M \frac{N_i}{K_{ij}-1}\sum_j^{K_{ij}}\sum_\ell (U_{ij\ell}-U_{ij})^2, \text{ and}$$

$$\hat{V}_2^2 = \frac{N-1}{N}\,B_2^2 + \frac{\bar{\bar{K}}-1}{\bar{\bar{K}}}\,W_2^2$$

$$\delta_2 \doteq \frac{B_2^2-\frac{W_2^2}{\bar{\bar{K}}}}{\hat{V}_2^2} \text{ (for large N).} \qquad (16)$$

The terms in (13) have an interpretation very similar to the corresponding terms in (7). The first term in brackets in (13) is the relative variance for a simple random sample of $n\bar{\bar{K}}$ people. The unbracketed middle term of (13) is a factor which represents the increase in the variance due to K_{ij} varying from HU to HU. δ_2 is the intraclass or intrahousehold correlation among people within households. Equivalently, δ_2 is a measure of the homogeneity among people in the same HU. Equating (12) and (13) yields the following expression for V_L^2:

$$V_L^2 = \left[\frac{V_2^2}{\bar{\bar{k}}}\right]\frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{k}}-1)] \qquad (17)$$

Substituting (17) in (7) and using (15) gives:

$$V_r^2 = \left[\frac{1-f}{mn\bar{\bar{K}}}\,V_2^2\right]\frac{V_2^2}{V_2^2}\,\frac{\hat{V}_L^2}{V_L^2}\,[1+\delta_L(\bar{n}-1)][1+\delta_2(\bar{\bar{k}}-1)] \qquad (18)$$

$$=\left[\frac{1-f}{mn\bar{\bar{K}}}\,\frac{1-R}{R\pi}\right]\mathrm{Def}_L\,\mathrm{Def}_2 \qquad (19)$$

where,

$$\mathrm{Def}_L = \frac{\hat{V}_L^2}{V_L^2}\,[1+\delta_L(\bar{n}-1)] \qquad (20)$$

and

$$\mathrm{Def}_2 = \frac{\hat{V}_2^2}{V_2^2}\,[1+\delta_2(\bar{\bar{k}}-1)] \qquad (21)$$

are the design effects for HU's within primary units and for people within HU's, respectively, and where $\pi = \bar{\bar{Y}}/K$ is the fraction of the population in the subgroup of interest. Assuming both design effects and their components can be approximated, the relative variance of r as given in (19) is finally in a desirable and usable form.

C. Estimating the Intraclass Correlations and Design Effects. We now turn to the estimation of the needed parameters. Data from the Current Population Survey (CPS), designed and conducted by the Census Bureau, were employed. The CPS [5] is a nationwide multi-stage sample survey conducted each month with a total sample size of about 56,000 designated HU's. Each election year, both presidential and nonpresidential, a supplement is added to the November CPS questionnaire which contains citizenship, registration, and voting questions. Although the November 1974 data were available, the November 1972 data were used in approximating the unknown parameters. The 1972 data were chosen for two important reasons. First of all, 1972 was a presidential election year, as was the election for which the survey was being designed.

Secondly, and quite fortunately, the 1972 sample consisted of a mixture of two sample designs. Half of the sample was the result of an older design for which $\bar{N}$=18, $\bar{n}$=6, and f_2=1/3. The other half of the sample stems from the CPS redesign and features $\bar{N}$=$\bar{n}$=4 and f_2=1. Having a reading for δ_L in both designs would indicate how δ_L (which is dependent upon $\bar{N}$) varies with changing $\bar{N}$. Only those counties which were self-representing (single counties or groups of counties are the primary sampling units in the CPS) in both designs were used in the estimation.

The total sample size for the study was approximately 20,000 HU's, about 10,000 sample HU's each for the old and new CPS sample designs. The analysis had several features. First of all, the counties in the analysis were placed, on the basis of geographic proximity, in one of four groups or universes and a fifth or combined universe which consisted of every county. The four groups consisted of counties in the Northeast, North Central, Southern and Western regions. Within each CPS sample design, the sample size was about 2,500 HU's in each of the first four universes. Second, since the only race designations collected in the 1972 CPS were White, Black, and Other, these three races, along with a fourth domain which included All races, were each used separately in the analysis. This resulted in 2x5x4=40 (number sample designs x number universes x number races) separate readings on the intraclass correlations and design effects. Unfortunately, due to small sample sizes for both the Black and Other racial subgroups, the readings obtained for these subgroups were considered highly suspect and consequently were of little use. Thus, the only reliable results were those obtained from the racial subgroups of White and All.

The usual consistent estimators for the unknown population parameters were employed in arriving at the following useful approximations for the parameters:

HU's

$$\bar{N}=4:\begin{cases}\delta_L = .166\\ \dfrac{\hat{V}_L^2}{V_L^2} = 1.05\\ \text{Def}_L = 1.05\ [1+.166(\bar{n}-1)]\end{cases} \quad (23)$$

$$\bar{N}=18:\begin{cases}\delta_L = .144\\ \dfrac{\hat{V}_L^2}{V_L^2} = 1.05\\ \text{Def}_L = 1.05\ [1+.144(\bar{n}-1)]\end{cases} \quad (24)$$

Persons

$$\bar{\bar{K}}=2:\begin{cases}\delta_2 = .627\\ \dfrac{\hat{V}_2^2}{V_2^2} = 1.15\\ \text{Def}_2 = 1.15\ [1+.627(\bar{\bar{K}}-1)]\end{cases} \quad (25)$$

Of course, the two underlying assumptions regarding the above conclusions are that (1) the minority and majority are fairly similar with respect to the above parameters, and (2) the counties in the actual survey are not unlike those in the CPS study. If one subscribes to the familiar model $\delta_L=a\bar{N}^b$ and uses the above results for $\bar{N}=4$ and $\bar{N}=18$, to solve for a and b, the obtained solution is,

$$\delta_L=(.1892)(\bar{N})^{-.0945} \quad (26)$$

which clearly demonstrates the dependence of δ_L upon $\bar{N}$. Likewise, δ_2 depends upon $\bar{\bar{K}}$, but since $\bar{\bar{K}}$ varied only slightly (about 2) among the jurisdictions in the actual Voting Rights Survey, equations (25) were considered valid for all jurisdictions (i.e., all $\bar{\bar{K}}$). Thus, the variance function (19) becomes:

$$V_r^2 = \left[\frac{1-f}{mn\bar{\bar{K}}}\ \frac{1-R}{R\ \pi}\right] 1.05[1+\delta_L(\bar{n}-1)]\ 1.15[1+.627(\bar{\bar{K}}-1)] \quad (27)$$

where δ_L depends upon $\bar{N}$ as discussed above. In the actual design of the survey, the value of π for a given jurisdiction was approximated by the 1970 census value and the value of R (obviously unknown) was taken to be the smaller of the 1972 overall voting participation rate for the entire jurisdiction (as given by Richard Scammon's "American Votes" [4] series) and the regional (in some cases national) minority of interest voting rate as estimated by the 1972 CPS. The values of π varied widely from .05 in some jurisdictions to a maximum of about .60, while the assumed minority voting rate generally satisfied $.20 \leq R \leq .40$.

III. SAMPLING FRAMES, COST CONSIDERATIONS AND OPTIMUM CLUSTER SIZES, ALLOCATION OF THE SAMPLE, AND DETERMINING SAMPLE/CENSUS STATUS FOR THE TOWN AND COUNTY JURISDICTIONS

This section will present several very closely related topics in the town and county jurisdictions. The various sampling frames frequently used by the Census Bureau to select general population samples will be described, as will the associated advantages, restrictions and costs for each frame. These costs, along with the already determined variance function, determine the approximate optimum cluster size in the various frames.

Also discussed will be the allocation of the sample to the various frames and strata. This section will conclude with some brief remarks concerning the determination of the sample versus census status of each jurisdiction.

A. Sampling Frames. There are three basic sampling frames which are used by the Census Bureau to select general population samples. A short description now follows for each of these three frames.

1. Old Construction Frames-1970 Census Detail Files. These are a group of files consisting of a detailed record for each, or a subset thereof, April 1970 housing unit. These files are a result of the 1970 census. The files that contain only a subset of the census units are the result of a sample and contain more detailed information for a given unit then does the complete tape. One large advantage of sampling from any of these files is the ease with which a high degree of stratification, based upon 1970 characteristics, is achieved. Of course, due to the movement of the population, the effectiveness of any stratification based upon 1970 characteristics decreases with time. Since units existing prior to April 1970 are referred to as old construction units, the above set of files will be referred to as old construction sampling frames.

2. New Construction Frame-Building Permits. Many counties require and maintain records of all newly constructed inhabitable structures in part or all of the county. These records generally take the form of building permits and contain the number of new HU's existing within the structure. Thus, with the aid of building permits, new HU's built in the permit issuing portions of a county can be sampled. Only a limited amount of stratification can be achieved, however, when sampling from building permits. Unfortunately, the permit issuing portion of a county may either be very small or nonexistent and thus, this building permit frame is often not available. Defining new construction units as those built since April 1970 clearly makes the building permit frame a new construction one.

3. Old and New Construction Frame-Area Maps. Another type of sampling that is widely used at the Bureau is area segmenting and sampling. The sampling frame used in area sampling is a land map showing the 1970 census count of HU's in small land areas. Each small land area (i.e., cluster) contains about twenty (i.e., $\bar{N}=20$) HU's, however, there is a fair amount of variation among these cluster sizes. These area segments are generally sampled with probability proportional to their size (i.e., 1970 HU count) and then subsampled as desired. The achievable degree of stratification is minimal with this type of area sampling, and further, as time goes on, the measures of size become poorer and poorer due to additions and losses of HU's. The advantage of the area frame is the ability to assign positive probabilities of selection to units built after the 1970 census, thus providing an alternative to sampling from building permits which are often unavailable. In addition, the area frame is also

used to sample old construction whenever census addresses from the old construction frame are poor. The area frame is obviously an old and new construction sampling frame.

B. Cost Considerations and Optimum Cluster Sizes. The determination of an appropriate cost model to be used in approximating optimum cluster sizes is often as important and as difficult as the derivation of the variance function. For this survey the importance of the variance function was probably greater than that of the cost function simply because of the strict reliability requirement. Since the emphasis in this paper is on the variance function and its detailed determination, we will sometimes be content with a fairly macroscopic discussion of some of the cost considerations.

In relative terms, it is generally more expensive to sample from the area frame than from either of the remaining two frames, which are each about equally expensive. Thus, it is desirable, from a cost standpoint, to use the 1970 detail files in conjunction with the building permit frame in the permit issuing portions of a given county. There is little choice but to use the area frame in the nonpermit issuing portions. The determination of the cluster sizes in the three frames will now be discussed.

1. Cluster Sizes in the New Construction Frame. It was decided to use the traditional permit new construction sample design of $\bar{N}=\bar{n}=4$. This type of clustering is frequently used at the Bureau and there was some advantage in being able to use established procedures. Also, a very rough cost analysis indicated this to be reasonably optimum. In addition, the new construction sample was generally a very small fraction of the overall sample, thus reducing the importance of optimality.

2. Cost Model for the Old Construction and Area Frames. The standard three term cost equation was developed and employed within each county jurisdiction. The cost model, derived separately for each of the old construction and area sampling frames within each county, took the following form:

$$c=c_1 m + c_2 \bar{n} m + c_3 \sqrt{mA}\ , \qquad (28)$$

where

c = total variable cost,

c_1 = cost per primary unit or cluster (includes cost of selecting, listing, and subsampling the clusters),

c_2 = cost per secondary unit or HU (includes cost of interviewing, processing, and within primary unit travel),

c_3 = cost per mile of travel between clusters (includes mileage and interviewer wages while traveling), and

A = county land area in square miles.

Without discussing the detailed computation of the actual cost coefficients (i.e., c_1, c_2, and c_3), the following observations are of extreme importance when determining the optimum cluster sizes within each sampling frame:

a) For each of the three frames the third term involving travel costs ($c_3\sqrt{Am}$) is negligible compared to the second term ($c_2\bar{n}m$) due to the small land areas A (often less than 500 square miles) generally encountered and due to the fairly large sample size $\bar{n}m$ (usually at least 500 sample HU's).

b) For the old construction frame the second term ($c_2\bar{n}m$) dominates the first term ($c_1 m$). This claim cannot be made for the area frame.

3. Optimum $\bar{N}$ and $\bar{n}$ in the Old Construction Frame. Assume the entire sample is to come from the old construction frame in a given county, subject to meeting the CV reliability requirement $\sqrt{V_r^2} = .10$, where V_r^2 is given by (27). The objective is to minimize the cost c in (28), while attaining this 10 percent CV. Though there is no control over $\bar{\bar{K}}(=\bar{\bar{k}})$, the cheapest combination of $\bar{N}$ and $\bar{n}(\leq\bar{N})$ can be selected. Although no mathematical solution exists for this particular problem, an iterative solution can easily be found as follows. Using (26) for δ_L as a function of $\bar{N}$, the only unknowns in V_r^2, as displayed in (27), for a given jurisdiction are m, $\bar{n}$, and $\bar{N}$. Specifying a given combination of $\bar{n}$ and $\bar{N}$, subject to $\bar{n}\leq\bar{N}$, one can solve for m using (27) and apply (28) to obtain the cost for this $\bar{n}$, $\bar{N}$, m combination. This procedure was followed for all reasonable combinations of $\bar{n}$ and $\bar{N}$ and the old construction sampling frame cost was recorded each time. As one would expect upon returning to the two comments immediately following (28), the winning combination in each county jurisdiction was $\bar{N}=\bar{n}=1$, or equivalently, a simple random sample of HU's.

4. Optimum $\bar{n}$ in the Area Frame. Assume the entire sample is to come from the area frame in a given county. Unlike the old construction sampling frame, it is not possible to select at will the value of $\bar{N}$ in the area frame. This is due to the nature of the area segmenting, in which the HU cluster sizes are variable and average about $\bar{N}=20$. This frame imposed restriction on $\bar{N}$ is, in some sense, similar to the real world imposed restriction on $\bar{\bar{K}}$, over which we have no control either. The area frame optimization procedure was similar to the old construction one, except only one value of $\bar{N}$ was considered, that value being 20. The cost for all combinations of $\bar{n}$ and m such that $V_r^2=.01$ and $\bar{n}\leq\bar{N}=20$ were determined. The cost efficient cluster size in each county jurisdiction was $\bar{n}=4$ HU's.

5. Optima in the Combined Sampling Frames. The optima just derived pertained to the old construction sampling frame (useful in the 100 percent permit issuing jurisdictions) and to the area frame (useful in the 0 percent permit issuing counties). About ¼ of the county jurisdictions were 100 percent permit issuing and a handful were 0 percent permit issuing. Thus, there were many partially permit issuing counties for which it was necessary to select a sample from each of the three frames. In such counties, it was decided to simply use the already determined optima in the various frames. That is, $\bar{N}=\bar{n}=4$ was used in the new construction frame, $\bar{N}=\bar{n}=1$ in the old construction frame, and $\bar{N}=20$, $\bar{n}=4$ in the area frame. Combining the individual sample frame optima to obtain an overall optima is permissible whenever the between cluster travel costs are relatively negligible (see Cochran [1], p. 289), as they are here.

C. Allocation of the Sample. The next step in the sample design was to efficiently allocate the sample to the various sampling frames or strata. When sampling from the 1970 detail tapes, the old construction frame was divided into two strata, those 1970 HU's with and without a minority of interest head. Thus, altogether there are four strata, the two old construction frame strata, the new construction stratum and the area stratum, to which the sample needed to be allocated. A variance function, similar to the earlier one but applicable to a stratified sample design will now be derived. The notation about to be introduced will be an obvious modification of the earlier notation with the first subscript (h) denoting the stratum rather than the cluster. For example:

Y_h = number of 18+ minority of interest citizens in stratum h, (h=1,2,3,4),

$R_h = \frac{X_h}{Y_h}$ = minority of interest voting rate in stratum h,

y'_h = usual unbiased estimator of Y_h based upon a sample of size n_h from N_h.

$r_s = \frac{\sum^4 x'_h}{\sum^4 y'_h}$ = stratified ratio estimator of R, and

Def_{Lh} = design effect for HU's within clusters in stratum h

$$= \left(\hat{V}^2_{Lh} \Big/ V^2_{Lh}\right)[1+\delta_{Lh}(\bar{n}_h - 1)]$$

$$= \begin{cases} 1.000 \text{ in the two old construction strata (h=1,2)} \\ 1.05\ [1+.166(4-1)] = 1.573 \text{ in the new construction stratum (h=3)} \\ 1.05\ [1+.143(4-1)] = 1.500 \text{ in the area stratum (h=4)} \end{cases}$$

The relative variance, $V^2_{r_s}$, of the stratified ratio estimator r_s is

$$V^2_{r_s} = \frac{Var(r_s)}{R^2} \doteq \frac{1}{X^2} \sum^4 Var(x'_h - R\, y'_h).$$

If the minority voting rate is assumed to be approximately the same in each stratum (probably a reasonable assumption), then $R_h = R$ (h=1,2,3,4) and we have

$$V^2_{r_s} \doteq \frac{1}{X^2} \sum^4 Y_h^2\, Var\left(\frac{x'_h - R_h\, y'_h}{Y_h}\right)$$

$$\doteq \frac{1}{X^2} \sum^4 Y_h^2\, Var(r_h) \doteq \frac{1}{X^2} \sum^4 X_h^2\, V^2_{r_h}.$$

The variance function for $V^2_{r_h}$ has already been derived and is given in (27). Using this result yields:

$$V^2_{r_s} \doteq \frac{1}{X^2} \sum^4 X_h^2 \left[\frac{1-f_h}{n_h \bar{\bar{K}}_h} \frac{1-R_h}{R_h \pi_h}\right] Def_{Lh} Def_{2h}$$

$$\doteq \frac{1-R}{RY^2} \sum^4 (1-f_h) \frac{Y_h N_h}{n_h} Def_{Lh} Def_{2h}$$

And finally, if one assumes $\bar{\bar{K}}_h = \bar{\bar{K}}$ (h=1,2,3,4), then we have:

$$V^2_{r_s} \doteq \left[\frac{1-R}{RY^2} Def_2\right] \sum^4 (1-f_h) \frac{Y_h N_h}{n_h} Def_{Lh}.$$

Since the HU costs in the four strata do not differ by more than a factor of two, a Neyman allocation is approximately optimal. Therefore, the sample was allocated to the four strata so that n_h was proportional to $\sqrt{Y_h N_h\, Def_{Lh}}$. In order to perform this allocation, estimates of Y_h and N_h (these are 1976 parameters and hence unknown) were needed. Based primarily upon the 5-year movement rates between 1965 and 1970 for each county, the known 1970 values of Y_h and N_h, the available estimates of new construction as well as a few other assumptions regarding the expected number of people moving into and out of an area, estimates of Y_h and N_h were made and used in the sample allocation.

D. Sample Vs. Census Jurisdictions. The final topic in this varied section will briefly discuss the determination of the sample and census jurisdictions. Costs and selection methods differ markedly between sample surveys and complete censuses. For example, an interviewed census HU will typically cost about $5.00 while the same HU selected by a sample survey might cost about $25.00. This would imply that whenever the sampling fraction $f = n/N$ is greater than .2, a census would be less expensive. Therefore, after determining the sample size and the corresponding sample survey cost for each town and county jurisdiction, and comparing this to the census cost, the sample and census jurisdictions were easily designated. As previously mentioned, in all 11 towns and in 19 of the 73 counties, it was cheaper to conduct a census.

IV. VARIANCE MODEL AND OPTIMA DETERMINATION IN THE STATE JURISDICTIONS

In 9, mostly southern, States, we were required to select a statewide sample in order to estimate the statewide minority of interest voting rate with a 10 percent CV. These 9 State jurisdictions included Arizona, Alaska, Alabama, Georgia, Louisiana, Mississippi, South Carolina, Texas, and Virginia. Arizona, which had 9 of its 14 counties designated as jurisdictions to be surveyed, was the only State among the 9 containing designated county jurisdictions. This section presents the derivation of the variance and cost functions that were extremely valuable in approximating the optimum combination of within PSU sample size, number of sample PSU's, PSU measure of size, and the within PSU cluster size, for each of the 9 States. Other aspects of the statewide sample designs are also discussed.

A. Variance Function. The first topic is the derivation of the all-important variance function. The goal, as has been the case throughout this paper, was to develop a variance model in terms of known or reasonably estimated parameters. In particular, it was also desired, at some point, to make use of the already determined within county variance model of section II.

The basic sampling plan is the following stratified multi-stage design. With the counties designated as the PSU's, stratify the PSU's, select a sample of PSU's from each stratum with replacement and with probability proportional to some measure of size, and subsample the chosen PSU's by first selecting clusters of HU's and then

subsampling the chosen clusters. The final two stages of selection that follow the first stage selection of PSU's is similar to the earlier within county sampling.

The notation to be used in deriving a variance function for this three-stage design is again an obvious modification of the original notation. Each of the original subscripts is to be shifted two places to the right. The first subscript (h) will now designate the stratum and the second subscript (p) will denote the PSU within the stratum. The third subscript (i) denotes the secondary unit (cluster), the fourth (j) denotes the third stage unit (HU), and the fifth (ℓ) denotes the individual people. For example, this new notation results in the following:

Y_h = number of 18+ minority of interest citizens in stratum h (h=1,2,...,H),

$R_h = \frac{X_h}{Y_h}$ = minority of interest voting rate in stratum h,

Y_{hp} = number of 18+ minority of interest citizens in PSU p of stratum h (p=1,2,...,T_h),

$R_{hp} = \frac{X_{hp}}{Y_{hp}}$ = minority of interest voting rate in PSU p of stratum h,

y'_{hp} = usual unbiased estimator of Y_{hp} based upon a sample of size n_{hp} HU's from the N_{hp} HU's in PSU p of stratum h,

where

H= number of strata in the State, and

T_h = number of PSU's in stratum h.

Also let:

t_h = number of sample PSU's in stratum h,

Z_{hp} = single-draw probabilities or normalized measures of size for PSU p of stratum h such that $\sum_p^{T_h} Z_{hp} = 1$,

y'_h = usual with replacement estimator of Y_h

$$= \sum_p^{T_h} \frac{y'_{hp}}{t_h Z_{hp}},$$

y' = usual stratified estimator of $Y = \sum^H y'_h$,

$r_M = \frac{x'}{y'}$ = multistage ratio estimator of R,

Def_{Lhp} = design effect for HU's within clusters in PSU p of stratum h, and

Def_{2hp} = design effect for people within HU's in PSU p of stratum h.

The relative variance of r_M, $V^2_{r_M}$, is first expressed as:

$$V^2_{r_M} = \frac{Var(r_M)}{R^2} \doteq \frac{Var(x'-R\,y')}{X^2} \doteq \frac{1}{X^2}\sum^H Var(x'_h - R y'_h) \quad (29)$$

Using Durbin's [2] (1953) well-known result concerning the variance of a multi-stage statistic, the general term of (29) can be expressed as:

$$Var(x'_h - R\,y'_h) = Var[E(x'_h - R\,y'_h | PSU's)] + E[Var(x'_h - R\,y'_h | PSU's)]$$

$$= \sum_p^{T_h} \frac{Z_{hp}}{t_h}\left[\frac{X_{hp}-RY_{hp}}{Z_{hp}} - (X_h - RY_h)\right]^2 + \sum_p^{T_h} \frac{1}{t_h Z_{hp}} Var(x'_{hp} - R y'_{hp} | PSU's). \quad (30)$$

The first term in (30) represents the familiar between-PSU variance and the second term the within PSU variance. To simplify the conditional within county variance $Var(x'_{hp} - R\,y'_{hp} | PSU's)$, the earlier results (14) and (18) are applied to the variate $U_{hpij\ell} = X_{hpij\ell} - RY_{hpij\ell}$. Ignoring the finite population correction (fpc) factor, this yields:

$$Var(x'_{hp} - R y'_{hp} | h,p) = K^2_{hp} \frac{\sum\sum\sum (U_{hpij\ell} - \bar{U}_{hp})^2}{(K_{hp}-1)\, n_{hp} \bar{\bar{K}}_{hp}} Def_{Lhp} Def_{2hp}. \quad (31)$$

Notice that in (31) it has been subtly assumed that the design effects for the variates $X_{hpij\ell} - R\,Y_{hpij\ell}$ and $X_{hpij\ell} - R_{hp} Y_{hpij\ell}$ are the same. This seems like a reasonable assumption, as design effects are usually fairly robust and these two variates are quite similar. Upon simplifying (31) we obtain

$$Var(x'_{hp} - R y'_{hp} | h,p) = \frac{N_{hp}}{n_{hp}}\left[X_{hp} + R^2 Y_{hp} - 2R\,X_{hp} - \frac{(X_{hp} - R\,Y_{hp})^2}{K_{hp}}\right] Def_{Lhp}\, Def_{2hp} \quad (32)$$

The variance function can now be assembled and expressed as

$$V^2_{r_M} = \sum_h^H \sum_p^{T_h} \frac{(X_{hp} - R\,Y_{hp})^2}{t_h Z_{hp} X^2} - \sum^H \frac{(X_h - R\,Y_h)^2}{t_h X^2} + \frac{1}{X^2}\sum_h^H \sum_p^{T_h} \frac{N_{hp}}{t_h Z_{hp} n_{hp}}\left[X_{hp} + R^2 Y_{hp} - 2RX_{hp} - \frac{(X_{hp} - RY_{hp})^2}{K_{hp}}\right] x\, Def_{Lhp} Def_{2hp}. \quad (33)$$

The first two terms in (33) is the simplified between-PSU relative variance of (30), with the second term explicitly showing the reduction in the total variance due to the stratification. Believe it or not, if one has an available computer, (33) is in a very usable form. X_{hp} can be estimated using 1972 county voting data from Scammon [4] and adjusting to account for the lower minority voting rates and the change in population between 1972 and 1976. N_{hp}, K_{hp}, and Y_{hp} can be estimated using 1970 census data and adjusting for the change in population between 1970 and 1976. Thus, the only unknowns in (33) are the formation of the strata, t_h, Z_{hp}, $\bar{N}_{hp}$, $\bar{n}_{hp}$, and n_{hp}.

For the moment, to aid in the search for the various optima, the following restrictions are placed upon our sample design:

1. Only one stratum will be formed and t PSU's will be selected with replacement from this single statewide stratum.
2. n_{hp} will be assumed constant for all PSU's and be denoted by W (workload).
3. $\bar{N}_{hp} \equiv \bar{N} = 20$ for all PSU's, primarily because it must equal 20 for any area sample.
4. $n_{hp} \equiv \bar{n}$ will be assumed constant for all PSU's thus $Def_L \equiv Def_{Lhp} = 1.05[1 + .143(\bar{n}-1)]$.
5. $Def_{2hp} \equiv Def_2$ is constant in each PSU.

Restrictions 1 and 2 above will later be lifted. Denoting the only stratum by h=1, the variance function (33) under these restrictions becomes,

$$V^2_{r_M} = \sum_p^{T_1} \frac{(X_{1p}-R\,Y_{1p})^2}{t\,Z_{1p}X^2} + \frac{Def_L Def_2}{X^2\,tW} \sum_p^{T_1} \frac{N_{1p}}{Z_{1p}}\left[X_{1p}+R^2Y_{1p}-2RX_{1p}\right] - \frac{Def_L Def_2}{X^2\,tW} \sum_p^{T_1} \frac{N_{1p}}{Z_{1p}}\left[\frac{(X_{1p}-RY_{1p})^2}{K_{1p}}\right]. \quad (34)$$

The unknowns for which the jointly optimum combination is desired have now been reduced to t, Z_{1p}, $\bar{n}(\leq 20)$, and W. Setting $V^2_{r_M}$ =.01 and specifying the set of basic probabilities or measures of size Z_{1p}, along with any two of t, $\bar{n}$, and W, will determine the remaining unspecified value uniquely.

The real innovation here is the attempt to find the optimal measures of size, i.e., the Z_{1p}'s. For various reasons, such as the desire for a selfweighting sample, these measures are generally taken to be proportional to the total number of HU's or the total population in a county. In addition, it is not very often that a survey is designed for the sole purpose of estimating one or two parameters, as was the case here. It is of interest to note the result obtained for $Z_{1p}=N_{1p}/N$ (i.e., probability proportional to the number of HU's) in (34). In this case $V^2_{r_M}$ simplifies to:

$$V^2_{r_M} \doteq \frac{N}{tX^2} \sum_p^{T_1} \frac{(X_{1p}-RY_{1p})^2}{N_{1p}} + \frac{(1-R)\,Def_L Def_2}{\left(\frac{tW}{\bar{n}}\right)\bar{n}\,\bar{\bar{K}}\,R\,\pi} \quad (35)$$

where a negligible term has been discarded. As seen from (19), apart from the fpc, the second term in (35) is simply the relative variance for the familiar two-stage cluster sample of size $tW/\bar{n}$ clusters and tW HU's selected from across the entire State, without regard to the county from which they arise.

To determine the optimum combination of t, Z_{1p}, $\bar{n}$, and W, a cost function is needed.

B. Cost Function. A brief description of the cost equations will now be given. The cost model is similar to the earlier one except for an additional term to account for the variable cost associated with the sample PSU's. The cost function for a State is given by:

$$C_M = C_{M1}(t) + C_{M2}\left(t\frac{W}{\bar{n}}\right) + C_{M3}(tW) + C_{M4}\,t\sqrt{\frac{W}{\bar{n}}}\,\bar{A}, \quad (36)$$

where

C_M = total variable cost for the State,

C_{M1} = cost per sample PSU (includes the cost of hiring and supervising the interviewer in a sample PSU),

C_{M2} = cost per cluster,

C_{M3} = cost per HU,

C_{M4} = travel cost between clusters in the same PSU, and

$\bar{A}$ = average county land area (square miles) in the State.

The four cost coefficients were computed separately for each State.

C. Determining the Optima. The method by which the optima were approximated will now be described. As before, no exact mathematical solution exists, however, computer assisted iterative optimization solutions over all possible reasonable combinations of the unknowns ($t, Z_{1p}, \bar{n}, W$) are easily found. Separately for each State, we specified the following 360 combinations of the probabilities Z_{1p}, the workload W, and the cluster size $\bar{n}$:

1. Six sets of probabilities, Z_{1p}:

$$\frac{P_{1p}}{P}, \frac{K_{1p}}{K}, \frac{N_{1p}}{N}, \frac{Y_{1p}}{Y}, \frac{\sqrt{Y_{1p}K_{1p}}}{\sum_q^{T_1}\sqrt{Y_{1q}K_{1q}}}, \frac{1}{T_1}$$

where P_{1p} = total population in PSU p, and

$P = \sum_p^{T_1} P_{1p}$ = total population in the State.

2. Ten workloads, W: 50, 100, 150, 200, 250, 300, 400, 450, 500
3. Six cluster sizes, $\bar{n}$: 1, 2, 3, 4, 5, 6.

The second and third sets of the Z_{1p} listed above are slight variations of the conventional measure P_{1p} (the first set).

The fourth set was investigated because one would expect it to identify areas of large numbers of minority. It was also hoped that the fifth set of the Z_{1p} would identify "pockets" of high minority density. This set of basic draw probabilities is probably the most interesting of the six sets and the intuition behind it was based upon the sample allocation formula as given in III.C. Actually, it was completely unknown as to how well this fifth set would ultimately perform. The sixth and final set of the Z_{1p} was tested only for curiosity purposes and consistently resulted in ridiculous optima, as expected.

Separately, for each specific combination of the Z_{1p}, W, and $\bar{n}$, expression (34) was used to solve the integer number of sample PSU's, t, necessary to satisfy $V^2_{r_M} \leq .01$. The cost of each specific combination of possible optima was then determined by using (36). The following table shows the resulting minimum cost combinations and other information for each State but Arizona, which, as mentioned earlier, was unique in that 9 of its 14 counties were already county jurisdictions. As the table shows, the measures of size Y_{1p} and $\sqrt{Y_{1p}K_{1p}}$ both performed quite well. Not shown in the table is the fact that for a given State, whenever Y_{1p} was the optimum measure of size, then $\sqrt{Y_{1p}K_{1p}}$ was never far behind, and conversely. Except for Texas, all optima shown in the table were actually used in the sample selection. For Texas, three sets (rows) of optima are listed, with the first and second sets corresponding only to the Black or the Spanish minorities, respectively. The third combination listed was the one used in Texas and was approximately optimal when considering both the Black and Spanish minorities. As a matter of fact, in Texas, additional sets of Z_{1p} were investigated

which were functions of both the Black Y_{1p} and the Spanish Y_{1p}. However, these special measures of size generally performed worse than the conventional P_{1p}, which was ultimately used. Since the optima were fairly flat, it was not uncommon to find that $\bar{n}$=3,5, or 6 (along with the appropriate combination of Z_{1p}, W_2 and t) was approximately optimal, along with $\bar{n}$=4. In these toss-up cases, the set of optima with $\bar{n}$=4 was chosen because of the advantages of using established sampling procedures in our three frames. Finally, the last column in the table indicates the amount of money that was saved by our optimization procedure over an alternative procedure which fixes the $Z_{1p} = P_{1p}/P$ (the conventional measures) and then optimizes.

D. Stratifying the PSU's. After determining the above statewide optima, the first two restrictions imposed by our model (34) were relaxed. Each State was stratified using approximately equal size strata and one PSU was selected per stratum using the measures of size determined optimal for the State. The workloads were then slightly adjusted to reflect the differing stratum sizes. Strata were formed on the basis of the percent minority and the minority median family income in the counties. In addition, there was frequently one stratum in each State that contained counties with virtually no minorities. Although it is not desirable to have these small minority PSU's in sample, it was felt safer to guarantee one and only one such PSU in sample rather than take a chance of selecting none, one, or more than one.

Even though more than modest gains were expected from the stratification, the sample sizes were not reduced to reflect this gain. This decision was based upon the fact that there were considerable approximations both in developing (34) and in estimating the many county totals used in the optimization. The gains associated with the complete stratification have not been estimated, however, the gains associated with the inclusion of our certainty PSU's only, were expected to reduce the 10 percent CV to about 9.6 percent in each State. Under this modified model that considers our certainty PSU's, the between-PSU variance as a percent of the total variance ranges from 10 to 25 percent across the 8 States.

E. Within PSU Sampling. For the sample counties in each State the optimal cluster size was shown to be $\bar{n}$=4. Thus, within each of the three sampling frames in each county, cluster sizes of n=4 were employed. The workload in each sample county was allocated to the three frames exactly as described earlier for the county jurisdictions.

V. ALTERNATIVE SAMPLE DESIGNS AND THE 1978 VOTING RIGHTS SURVEY

This final section includes a brief discussion of the research into alternative sample designs that is currently taking place and of the upcoming 1978 Voting Rights Survey.

A. Alternative Sample Designs. The tendency for people to overreport voting and the resulting bias is a common problem in survey work. Although the 1976 sample design did not address this unfortunate phenomenon, it is planned to reduce the over-reporting bias, where possible, by ratio estimating to the actual overall number of votes cast as given by the jurisdictions themselves. In addition, the Bureau has begun research concerning two alternative sample designs that are expected to reduce the over-reporting bias at an affordable price.

The first alternative is a dual-frame sampling scheme. The two sampling frames in this scheme are (1) the usual Bureau frames described throughout this paper, and (2) county registration lists. A sample is drawn from each frame and a combined dual-frame estimator is employed. For a given amount of money, it is unknown as to whether or not the mean squared error of the dual-frame estimator is less than that of the conventional sample design estimator. Research is continuing in 12 county jurisdictions in an attempt to answer this question.

The second alternative sample design is a double sampling records check approach. In this design, the usual household survey is conducted and a subsample of the surveyed households is then selected. The voting responses for the persons in these subsampled households are then checked against voter and registration lists and an estimator reflecting the observed over-reporting in the subsample is formed. Again, our research seeks to determine the cost effectiveness of this double sampling scheme.

B. 1978 Voting Rights Survey. The Bureau conducted the 1976 Voting Rights survey in the 93 jurisdictions and the research discussed above for about $5,000,000. The 1976 survey, however, is small in both price and the number of covered jurisdictions compared to the 1978 Voting Rights Survey currently being planned. For the 1978 survey, the Bureau has been directed to treat each individual county in the nine States as a jurisdiction in its own right. In addition, the town and county jurisdictions covered in the 1976 survey are to be retained in 1978. Thus, the Bureau is expected to be given about $40,000,000 to conduct sample surveys or censuses in about 950 town and county jurisdictions in November 1978.

The innovative sample design strategy presently being planned for the 1978 survey is highly analytic in nature. We are attempting to divorce ourselves from the relatively artificial 10 percent CV reliability requirement concept and design the survey with the analyst and decision maker in mind. The power function is the key concept in our unique design. As of this writing, it is felt one of the best ways to spend the $40,000,000 is to design the 1978 survey so that in each sample jurisdiction, the probability of concluding the White voter participation rate is more than 3 percentage points higher than the minority voter participation rate (versus concluding the difference is exactly 3 percentage points), is equal to .10 when the true difference is 3 percentage points (a type I error), and is equal to .90 when the actual difference is 10 percentage points (a correct conclusion). In addition, the budget for the 1978 survey includes funds for a 100 percent voting records check, thus eliminating the over-reporting bias.

TABLE OF STATE OPTIMA

State	Minority of interest	π	R	Z_{1p} proportional to:	W	$\bar{n}$	t	Total sample size =tW	Total 4/ sample size if statewide SRS	Dollars 5/ saved over $Z_{1p} = \frac{P_{1p}}{P}$
Alaska	Native Alaskan	.13	.43	$\sqrt{Y_{1p}K_{1p}}$	150	4	11	1650	1050	$20,000
Alabama	Black	.23	.43	Y_{1p}	100	4	11	1100	600	$ 7,000
Georgia	Black	.22	.38	$\sqrt{Y_{1p}K_{1p}}$	100	4	13	1300	750	$ 8,000
Louisiana	Black	.26	.43	Y_{1p}	100	4	9	900	500	$ 3,000
Mississippi	Black	.31	.43	Y_{1p}	100	4	9	900	450	$ 2,000
So. Carolina	Black	.26	.39	Y_{1p}	100	4	11	1100	600	$ 2,000
Texas	Black 1/	.11	.43	$\sqrt{Y_{1p}K_{1p}}$	150	4	12	1800	1250	$15,000
	Spanish Heritage 2/	.14	.43	$\sqrt{Y_{1p}K_{1p}}$	100	4	13	1300	1050	$20,000
	Black, Spanish Heritage 3/	-	-	P_{1p}	100	4	21	2100	1250	0
Virginia	Black	.16	.43	$\sqrt{Y_{1p}K_{1p}}$	100	4	12	1200	800	$11,000

1/ Considers Black only, ignores Spanish Heritage. This design for Black would yield an unacceptable 24 percent CV for Spanish Heritage.
2/ Considers Spanish Heritage only, ignores Blacks.
3/ This design jointly yields a 10 percent CV for Blacks and a 9.2 percent CV for Spanish.
4/ For comparison with tW, this column gives the sample size for a statewide simple random sample (SRS)
5/ This column gives the savings over the conventional design using probability proportional to total population, i.e., $Z_{1p} = P_{1p}/_{P}$.

REMINDER

π = fraction minority of interest
R = minority voting rate
Z_{1p} = single draw probabilities
W = Within PSU sample size
$\bar{n}$ = average HU cluster size
t = number sample PSU's

REFERENCES

1. W. G. Cochran. Sampling Techniques. 2nd ed. New York: Wiley and Sons, 1963.
2. J. Durbin. Some Results in Sampling When the Units are Selected with Unequal Probabilities. Journal of the Royal Statistical Society, Series B, (1953), Vol. 15, pp. 262-269.
3. M. H. Hansen, W. N. Hurwitz and W. G. Madow. Sample Survey Methods and Theory, Vol. 1, Methods and Applications. 1st ed. New York: Wiley and Sons, 1953.
4. R. Scammon. American Votes 10. 10th ed. Washington, D. C.: Congressional Quarterly, 1972.
5. M. Thompson and G. Shapiro. The Current Population Survey: An Overview. Annals of Economic and Social Measurement, (1973), Vol. 2, No. 2.

ACKNOWLEDGMENTS

The author would like to thank Bob Jewett and Duane Hybertson for their excellent computer programming and Edith Oechsler for her careful typing; all of the U. S. Census Bureau.

COVERAGE IMPROVEMENT IN THE ANNUAL HOUSING SURVEY

Irene C. Montie and Dennis J. Schwanz, U.S. Bureau of the Census

ACKNOWLEDGEMENT

The intent of this paper is to describe some methodologies used to reduce undercoverage bias in the Annual Housing Survey, National Sample.

Identification of deficiencies in the sampling frames and development of methodologies to close the gaps have been a joint effort within Statistical Methods Division. The work of the various statisticians are so commingled that individual acknowledgement would mask the synergistic process. Therefore, in this paper footnotes are limited to peripheral publications. The reader is directed to Census Bureau planning, procedural, and results memoranda from which this effort is derived. However, some recognition must be given to Dave Bateman and his sample design staff, Len Baer and his sampling systems staff, the computer programming staff under Dave Diskin, the staffs of sampling procedures branch and special surveys branch, and of course the Division Chief, Charlie Jones, for his relentless encouragement in carrying out the coverage improvement procedures. Special notice must also be given to the Field Division staff, who implemented the procedures, with precision and concern, and to Arthur F. Young Chief of Housing Division, Elmo Beach and his staff, and the staff of Demographic Surveys Division for their knowledgeable assistance, critique, and support of the entire undertaking.

I. SURVEY BACKGROUND

The Annual Housing Survey (AHS-National Sample) is a sample survey conducted annually by the Bureau of the Census for the Department of Housing and Urban Development to obtain national and regional estimates of the size and composition of the housing inventory in the United States. The series estimates year to year changes in the inventory due to losses and new construction (including mobile home placements), and provides characteristics of the total inventory.

The survey was first conducted in 1973. At that time approximately 59,300 sample units were contacted. The 1974 sample included 1,358 additional units to represent new construction built since the 1973 survey. This updating of new construction has been continued on an annual basis. In addition, in 1974 the sample in rural areas was doubled (an increase of 15,500 units) to provide for greater precision in measuring certain characteristics of housing in rural areas. Each year, demolished units and other types of nonexistent units have been deleted from the sample, thus partially offsetting the increase from new construction. At present the sample consists of 81,850 units.

II. PURPOSE AND SCOPE OF THIS REPORT

This report is principally concerned with nonsampling errors related to undercoverage in the Annual Housing Survey (AHS-National Sample). In comparing the first year results to independently derived estimates it became apparent that certain types of units, such as mobile homes, were not adequately represented in the sample. The types of omissions had been generally recognized, but their magnitude and impact on components of the inventory had not been fully recognized. In particular, for mobile homes the undercoverage was compounded by census misses and definitional differences in the basic sampling frame.

The purpose of this report is to describe the types of undercoverage, the methodology for representing undercovered units in the sample, and their effect on the undercoverage bias. These topics are discussed in sections IV - VI below; summary and conclusions appear in section VII.

III. AHS SAMPLE DESIGN AND ESTIMATION PROCEDURES

A. Sample Design

The AHS is a multi-stage cluster sample of about 82,000 units spread over 461 PSU's, comprising 923 counties and independent cities. Of the 461 PSU's, 156 were included in sample with certainty; these are referred to as self-representing. The remaining PSU's were grouped into strata and a sample of PSU's was selected from each stratum. This resulted in an additional 305 PSU's, which are referred to as non-self-representing.

Within each sample PSU, a sample of units from the 1970 Decennial Census listings was selected. This was accomplished in several stages. First, a sample of census enumeration districts (ED's) was selected. The next stage consisted of the formation and selection of clusters of housing units (HU's) within each sample ED, where the selection method was dependent on the type of ED. There are two types of sample ED's - Address and Area. Address ED's are those in which building permits are authorized for new construction and at least 90 percent of the 1970 census addresses were listed with house number and street address. In these ED's a compact cluster of an expected four units was selected from the 1970 census address listings.

Area ED's are those that do not meet one or both of the Address ED criteria. These ED's were divided into small land areas referred to as area segments. Each area segment selected for AHS was canvassed and all units (both 1970 census units and units built after the census) were listed. A systematic sample was then selected from this listing for AHS; this resulted in a noncompact cluster of an expected four units in each area segment.

In addition, a sample of new construction building permits was selected within each sample PSU to represent units built after the census. These are called permit segments. Finally, as a result of a 1970 census evaluation study, a sample of units missed in the census was also included; these are referred to as CEN-SUP segments.

B. Estimation Procedure

The estimation procedure, utilized for AHS in 1973-1975, employed three stages of ratio estimation. The first stage was employed for sample units from NSR PSU's only and was designed to reduce the between-PSU component of variance, due to the sampling of PSU's.

The second stage ratio estimation, which is very relevant to the undercoverage problem, was only

employed for units built in April 1970, or later (new construction units). This procedure was designed to adjust the AHS sample estimates of new construction to independently derived new construction estimates that were considered to be the best estimates available. These estimates were derived from the Survey of Construction (SOC), a survey of building permits conducted monthly by the Census Bureau (for conventional new construction), and from mobile home shipments reported by the Mobile Home Manufacturers Association (for new mobile homes). This adjustment was necessary to correct for the undercoverage biases in AHS with respect to new construction.

The third stage ratio estimation was employed for all sample units. It was designed to adjust the AHS sample estimates to independently derived estimates for four types of vacant units and 24 residence-tenure-race of head-sex of head categories for occupied HU's. These estimates were derived from the Housing Vacancy Survey (HVS), a quarterly vacancy survey conducted by the Bureau, and the Current Population Survey (CPS), a monthly unemployment survey also conducted by the Bureau.

IV. SOURCE AND TYPE OF UNDERCOVERAGE

As noted in the previous section, there are two types of Enumeration Districts (ED's); i.e., those for which permits are issued for new construction (permit-issuing areas) and those for which permits are not required for new construction (nonpermit areas). This paper is concerned with undercoverage in address segments, which are located in permit-issuing ED's, and in permit segments, which are used to represent new construction in these ED's.

The frames used for selecting the sample in address segment areas have certain deficiencies which, in total, represent something less than 2 percent of the universe (about 1,080,000 units, of which about 959,000 are eligible to be counted in the housing inventory). However, the undercoverage is disproportionately concentrated in certain types of units. These units are described below, along with estimates of their undercoverage.

One source of undercoverage bias is in units constructed since the census. For AHS, new construction is defined as units created on the site, including prefabricated housing, and occupied new mobile home placements. Prefabricated housing is represented in address segment ED's through permit segments. However, units completed after the census for which permits were issued before January 1, 1970, are not included in the sampling frame. These are referred to as permit lag units and are estimated at about 598,000 units.

The other type of new construction consists of occupied mobile home placements for which the undercoverage bias is estimated at 294,000 units.[1/] These units may be located in mobile home parks or on individual lots at large. Some of these parks have been created since the census; others existed prior to 1970 but were either missed in the census or unreported due to definitional differences.[2/] There is also undercoverage of mobile homes that were manufactured prior to 1970.

Another source of coverage loss is nonresidential units that have been converted to residential use since the census. The permit universe consists of permits for residential new construction only; it does not include permits for alterations to existing structures. Although these conversions are a small component of the housing inventory, they have unique characteristics that may not be fully represented in the independent estimates used in the third stage sample adjustment and therefore, contribute to biases in the characteristics of the total inventory.

Houses that have been moved into address segments since the census are also undercovered. They have no chance of selection at the census address nor at the new address, unless they replace housing that existed at the new address at the time of the census. The estimate of this undercoverage is 50,000 units.

Procedures have been developed to represent all of these types of units in the AHS National sample. These coverage improvement procedures are described in section V.

V. COVERAGE IMPROVEMENT SAMPLE DESIGN AND IMPLEMENTATION

Four coverage improvement procedures were developed to reduce undercoverage bias of the types of units described in section IV. The design and implementation of the samples are discussed in this section; survey results appear in section VI.

A. Permit Lag

The permit lag sample provides coverage of new construction for which permits were issued prior to January 1, 1970 but construction was completed after the census.

1. Sample Design

The permit lag sampling frame was created from the Survey of Construction (SOC), a survey of authorized building permits conducted monthly by the Bureau of the Census to determine the rate at which these authorized units are started and completed. Between 1964 and 1973, SOC was conducted in a 122-PSU design, which was a subset of the CPS 449-PSU design. Within each of these PSU's, a sample of permits authorized each month was selected from each of the sample permit-issuing places. A three stage sample selection was used which resulted in an overall probability of selection of 1-in-100 for each sample permit.

For each permit in SOC, the month construction started and the month it was completed were determined. From this a sampling frame was created which consisted of permits for residential structures that had been authorized before January 1970 but were completed after the 1970 census (i.e., April 1970). The AHS permit lag sample was selected from this frame. However, some of these sample permits were in PSU's which were not in the AHS sample design or in any other sample design. It was decided to drop these permits from the sampling frame since interviewing units in these PSU's would not be cost effective. To compensate for these units, the weights associated with the remaining sample units were adjusted by the following factor:

$$\frac{\text{Wt'ed. HU's in non-AHS PSU's} + \text{Wt'ed. HU's in AHS PSUs}}{\text{Weighted HU's in AHS PSU's}}$$

For cost efficiency reasons, it was decided that the ultimate sampling unit for the permit lag sample should be a compact cluster of about four units. Thus, the units for each permit in the frame were divided into clusters of about four units.

Each permit in the frame had a measure of size which was the weighted number of HU's represented by the particular permit. Prior to selecting a sample of the clusters, this permit measure of size was transformed into a cluster measure of size according to the following formula:

$$\text{Measure of size of cluster } j = \frac{M_i}{(K_i)N_{ij}} \quad 3/$$

where: M_i is the measure of size for the i^{th} permit.

K_i is the number of clusters associated with the i^{th} permit.

N_{ij} is the size of the j^{th} cluster in the i^{th} permit.

Prior to sample selection, the clusters of this sample frame were stratified according to the following variables:

1. Size of structure
   a. 1 unit
   b. 2-3 units
   c. 4-5 units
   d. 6-7 units
   e. 8-9 units
   f. 10-16 units
   g. 17-49 units
   h. 50-99 units
   i. 100-199 units
   j. 200 or more units
2. Region
3. SMSA/Non-SMSA
4. PSU Number
5. Permit Number
6. Cluster Number

This stratification was employed to insure a representative sample of these types of units by size of structure, region, SMSA/Non-SMSA, etc.

Since all of the Bureau's recurring surveys (i.e., CPS, AHS, the National Crime Survey [NCS], and the Health Interview Survey [HIS] fail to properly represent these permits, representative national samples of clusters necessary for the rest of the decade were selected for each of these surveys. This included one sample for AHS, thirteen samples for CPS, six samples for NCS, eight samples for HIS, and two samples to be held in reserve for future surveys. The clusters were selected with probability proportionate to the cluster's measure of size at a rate of 1-in-47. The selected clusters or hits were assigned to each of these samples according to the following scheme:

Hit 1 : AHS
Hits 2-14 : CPS (samples A36-A48)
Hits 15-16: Reserve samples
Hits 17-19: NCS (samples J03, 05, 07)
Hits 20-27: HIS (samples Y77-Y84)
Hit 28 : AHS
Hits 29-41: CPS (samples A36-A48)
Hits 42-44: NCS (samples J04, 06, 08)
Hits 45-46: Reserve samples
Hits 47-54: HIS (Y77-Y84)
Hit 55 : AHS
Hits 56-58: NCS (samples J03, 05, 07)
Hits 59-71: CPS (samples C20-C32)
Hits 72-73: Reserve samples
Hits 74-76: NCS (J04, 06, 08)
Hits 77-84: HIS (Y77-Y84)

The assignment order presented in the above scheme was repeated for every 84 hits, which means that 3 out of every 84 selected clusters were assigned to AHS.

2. Systems and Procedures

As indicated above, the permit lag universe was developed from a computer listing of 12,920 permits issued during the years 1967-1969 in the sample PSU's. The permit issuing date and the date construction was completed appeared on the list for each unit. Thus the universe was created by stripping off addresses of all structures that were completed after April 1, 1970. A sample of 1,386 units was selected for the AHS national sample.

The selected units were clustered by geographic location into 438 segments of size 1-5. A total of 1,386 units were assigned for interview during the regular AHS interview period (roughly September - November 1976).

Some overlap between the permit universe and census addresses was discovered at time of AHS interview. This occurred, in part, because the reported date of completion for multi-unit structures was the date when more than half of the units were completed. Thus some of the units were completed earlier and could have been reported in the census. In these situations the basic address, and all units at it were eligible for inclusion in the AHS sample. In the case of single-unit structures the census enumerator could have considered construction sufficiently complete to report such units as vacant. (Some subjectivity entered into the determination of vacancy status.)

Overlap could also occur between the permit lag universe and the regular permit universe or the CEN-SUP sample, for methodological reasons or due to permit issuing practices. For example, all units at a sampled permit address are listed, regardless of the number of structures involved. However, separate permits may have been issued for each structure and, depending upon the timing, subsequent permits might not be discovered.

In the case of overlap with CEN-SUP, that sample was developed after the census and may have included some permit lag units. Since CEN-SUP is a sample, not a universe, and the PSU's in the permit lag universe are a subset of the PSU's for which CEN-SUP was developed, a complete unduplication cannot be accomplished.

The overlap among the various universes is expected to be small. However, it is presently under investigation. In addition, some procedural controls are imposed to correct the overlap. For example, the interviewer is told the number of units for which the permit is issued. If more units are found than expected, a check is made to determine if this is the result of overlapping frames, or due to permit problems such as overbuilding or underreporting on the permit. Adjustments in the sample estimates are made as a result of duplication discovered through procedural controls.

B. Woodall Sample

This sample was selected from a universe of mobile home parks obtained from a commercial list. The list was updated each year through 1974, when the commercial operation was terminated. Thus the Woodall sample provides coverage of mobile homes located in parks created after the census and through calendar year 1974. Parks that were begun before 1970, but completed after the census, also were included. (Mobile home parks and other special places are not included in the Permit Lag sample since they are not sampled from permits.)

1. Sample Design

This sample was designed to provide coverage of mobile homes located in parks which were created after the 1970 census. Since the sample was limited to address ED's, it was necessary to unduplicate these places from area segment ED's. In addition a check needed to be made against the Census listings for places reported as created through 1972 in case any part of these places existed at the time of the Census. To do this it was necessary to determine (as described in paragraph 2) the ED in which each park was located. The unduplication and matching procedures were costly and time consuming. In order to reduce costs and preparatory time, it was decided to implement this procedure in the 266-PSU design (the representative national sample of PSU's which is a subset of the AHS design). The savings in cost and time were considered sufficiently important to outweigh any increase in the between PSU variance component resulting from this design. Therefore, the Woodall sampling frame consisted of the mobile home parks on the Woodall commercial listing which (1) were identified as having been created after the 1970 census and (2) were located in an address ED in the 266-PSU design.

Since it was decided to employ noncompact clusters of size four for this procedure, similar to what is done for other mobile home parks in AHS, the measure of size associated with each park in the Woodall sample frame was equal to the following:

$$\frac{\text{Number of sites in park}}{4}$$

Prior to sample selection, the mobile home parks were stratified according to the following variables:

1. Region 2. SMSA/Non-SMSA 3. SR/NSR

This stratified sampling frame of mobile home parks was them sampled with probability proportionate to the park's measure of size such that the overall probability of selection for each hit, or sample cluster, was 1-in-1366. This resulted in 30 sample mobile home parks from which 31 noncompact clusters of 4 mobile home sites were selected for the AHS Woodall samples. The procedure for selecting the sample units appears below.

2. Systems and Procedures

The full universe consists of 794 parks that were not available for listing at the time of the census. The universe was created by determining the geographic location of each park on the commercial list and allocating the parks to the appropriate census ED. Then the ED's were identified as area or address segment ED's, according to their permit issuing status and certain other criteria related to the adequacy of addresses in the ED.[4/] Parks in area segment ED's were dropped from the universe because they had a chance of selection in the AHS sample through area segments.

Since a mobile home park would have a chance of selection in the AHS sample if even one unit was occupied at the time of the census, an unduplication procedure was mounted. The address of each park located in an address segment ED and completed before January 1, 1973 was matched against the census listings. Any parks listed in the census, as either a regular address or a special place address, were dropped from the Woodall universe. The January 1973 cutoff was used because it was felt that a park, for which construction had begun before the census, would be completed by that date. Because vacant mobile home sites were not reported in the census, a review was made of parks that first appeared on the commercial list in 1969. These were processed as described above, and those not found in the census were included in the Woodall universe.

In order to avoid clustering, interviewers listed all sites (occupied or vacant) at the selected parks and a non-clustered sample of approximately 4 sites was selected from the listings. A total of 119 sites were assigned for AHS interview.

C. Windshield Sample

The Windshield sample was used to supplement the Woodall sample. It was originally conceived as a source for providing coverage of mobile home parks created after the termination of the Woodall operation; i.e., after January 1, 1975. However, some preliminary investigation indicated that the Windshield sample had the potential for improving undercoverage bias of parks missed or otherwise unreported in the census and in the Woodall sample. Thus, the scope of the Windshield sample was broadened to provide for this additional coverage.

1. Sample Design

The Windshield sample design was a two stage sample selection procedure implemented in the entire AHS 461 PSU design. The first stage consisted of selecting about 150 tracts within these PSU's. It was decided to select tracts[5/] since they were small enough to be canvassed at relatively little cost and time, but were large enough to yield a significant payoff in terms of locating missing mobile home parks. One problem with using tracts as the area to be canvassed is that the sample was supposed to represent missing mobile home parks in address ED's only; but the sample tracts could contain some area ED's. This problem was resolved by eliminating all parks found to be in area ED's. The identification was made after the tracts had been canvassed, because it was more efficient than unduplicating the area ED's before canvassing the tracts. One-hundred and fifty tracts were selected because it was felt that this was the maximum number of tracts that could be canvassed, taking into consideration the time and cost constraints. Although

this was not necessarily the optimum number of tracts, it was felt that canvassing this number of tracts would result in a relatively reliable estimate of mobile homes in missing mobile home parks.

The 150 tracts were selected from a file, created from the 1970 census fourth count tape, that contained a record for each tract in the 461 PSU design. A measure of size (M_i'), equal to the total number of mobile homes in the tract as reported in the 1970 census was assigned to each tract.[6/] Even though the number of 1970 census mobile homes may not necessarily have been highly correlated with mobile homes in missing mobile home parks, it was felt that this measure of size was the best available for selecting the sample tracts. The measure of size was then adjusted by the inverse of the probability of selecting the PSU in which the tract was located, to reflect the sampling of NSR PSU's. The adjusted measure of size (M_i) was then used in the selection of sample tracts. This tract file was stratified, or sorted, by the following variables:

1. Region 2. SR/NSR PSU 3. M_i

The sample of tracts was then selected with probability proportionate to M_i using the following sampling fraction: $\frac{150}{M}$ (where M equals the sum of M_i's across all of the tracts in the 461 PSU's.)

The 150 selected tracts were then canvassed, as described in the next section. Mobile home parks identified in the canvassing operation that were found to be in area ED's, enumerated in the 1970 census, or duplicated on the Woodall list, were deleted from the Windshield sample.

The second stage procedure was the selection of noncompact clusters of size four (mobile home sites) within the remaining mobile home parks. Prior to this sample selection, the parks were sorted into two types - census misses (parks in existence in April 1970 which were not enumerated in the census) and Woodall misses (parks built after April 1970 which were not on the Woodall list). The second stage selection was implemented independently within each type of park. The noncompact clusters of size four were then sampled with equal probability within each park using the following sampling fraction for each sample tract:

$$\frac{1}{1366} \times \frac{M}{150\ M_i}$$

This within-tract sampling fraction was employed so that each noncompact cluster of four would have the same overall probability of selection, 1-in-1366, as the other AHS sample units (i.e., this sampling fraction was used to preserve, as much as possible, the self-weighting aspects of the AHS sample design).

2. Systems and Procedures

Census interviewers canvassed each of the 150 tracts selected in the Windshield sample. In order to reduce costs, all major roads were physically canvassed, as were any areas where signs indicated the location of a mobile home park, but inquiry was made in areas where parks were not likely to be located; e.g., in high cost housing projects.

A form was filled for each park discovered. This provided identification information and the size of the park. Through a matching operation the parks were unduplicated from the Woodall universe and from the census. This resulted in 85 parks, which were subsampled at a rate computed separately for each tract. A sample of 24 parks was selected, from which 29 segments were created. (Double hits occurred in some large parks.) In order to avoid clustering, a sample of units was selected across each park. A total of 118 units were assigned for AHS interview.

D. Successor Check

The successor check provides coverage of three types of units that would not have been reported in the census at their present location.

The first type is mobile homes at large (not located in parks) that were either placed on the present site since the census or were vacant at the time of the census. (Vacant mobile homes were not reported in the census even when they were affixed to a permanent foundation.) The second type is houses that were moved to the present site since the census. Finally, the successor check provides coverage of units in structures that were converted to residential use since the census. These three types of units are referred to as inscope successors.

1. Sample Design

Unlike the Permit Lag or Woodall coverage improvements, a universe (or sample-based) listing of these types of units, from which a representative sample could be selected, did not exist. Thus, it was decided to use a successor check procedure. This is a listing procedure that has been used previously by the Census Bureau (e.g., it was used in the spring of 1976 for the Survey of Income and Education and was used for CPS and HIS throughout the 1960 decade). The successor check procedure is described in more detail in the next section.

Briefly, it involves listing a string of k structures in a predetermined order. The string begins with an AHS sample unit and is bounded by the kth residential structure that existed in 1970. Inscope units are identified along the string, between these two structures.

Since the check was related to the AHS sample, the only sample design questions that needed to be resolved for this coverage improvement procedure were the size of the string (k) and how many strings should be listed (i.e., the number of AHS sample addresses from which a listing should be started).

The 1970 Components of Inventory Change Survey (CINCH) showed that between April 1960 and October 1970 there were about 743,000 of these types of units added to the inventory. This represented about 1 percent of the inventory in a 10 3/4 year time period; therefore, it was assumed that these units added since April 1970 represented about .6 percent of the total inventory. Since these units represented such a small fraction of the total inventory, it was assumed

to be unlikely that more than one inscope structure would be found in a string. Therefore, the intraclass correlation between inscope or missed structures would not depend on the string length which implied, in terms of variance constraints, that the string size should be as large as reasonable. This was also true to a certain extent, in terms of cost considerations. The cost per inscope unit decreases as the size of the string increases since the expected number of inscope structures listed per string also increases. However, the Bureau's field personnel felt that after a certain string size there would be a large incremental cost increase due to added complexity, travel, more supervisory referrals, etc. Although they did not know the exact size at which this increased cost would be incurred, it was speculated that this would happen for a string size of 12 or more. Even though it was not optimal, a string size of 8 was selected as a compromise, to minimize the risk of incurring this additional cost increase since the coverage improvement budget was very tight and would not allow for this additional cost.

Given the string size of eight, the number of strings was determined by equating this to an optimal allocation determination for a stratified sample involving two strata. The first stratum was the universe represented by the successor check units and the other stratum was the universe represented by the rest of the AHS sample units. This optimal allocation formula is given as follows:

$$\frac{n_{SC}}{n} = \frac{N_{SC}\, S_{SC} / \sqrt{C_{SC}}}{N_{AHS}\, S_{AHS} / \sqrt{C_{AHS}} + N_{SC}\, S_{SC} / \sqrt{C_{SC}}}$$

where:

N_{SC} = the number of units in the successor check universe.

N_{AHS} = the number of units in the regular AHS universe.

C_{SC} = cost per unit for the successor check universe (for a string of 8, this cost equalled $306.25).

C_{AHS} = cost per unit for the regular AHS universe (C_{AHS} = $24.50).

S^2_{SC} = the unit variance for the successor check universe.

S^2_{AHS} = the unit variance for the regular AHS universe.

We know that:

$N_{AHS} = (1-P_{SC})\, N$

$N_{SC} = P_{SC} N$

where P_{SC} = proportion of the total universe represented by the successor check universe.

$n = n_{SC} + n_{AHS}$

$= n_{SC} + 1.57\, n_{AHS(oc)}$

(where $n_{AHS(oc)}$ is the AHS sample size for units from address segments.)

Inverting the allocation formula and making the above substitution produced the following result:

$$\frac{n_{AHS(oc)}}{n_{SC}} = \frac{(1-P_{SC})}{(1.57)P_{SC}} \frac{S_{AHS}}{S_{SC}} \sqrt{C_{SC}/C_{AHS}}$$

Since the number of units in each address segment is two and the string size is eight, then:

$n_{AHS(oc)}$ = 2 (Number of address segments)

$n_{SC} = \frac{8\ Psc}{1-Psc}$ (Number of successor check strings)

Substituting the above into the allocation formula produces the following:

$$\frac{\text{Number of address segments}}{\text{Number of successor check strings}} = \frac{8}{2(1.57)} \frac{S_{AHS}}{S_{SC}} \sqrt{C_{SC}/C_{AHS}}$$

Since $\frac{S_{AHS}}{S_{SC}}$ did not vary greatly, the optimal ratio of address segments to the number of successor check strings for a string size of eight was determined by the square root of the ratio of the costs. This optimum subsampling rate was about twelve. In other words, one-twelfth of the AHS address segments (about 1,500) would be used as the starting points for the successor check strings. Since the AHS sample had been divided into six panels, each of which was a representative national sample, a systematic half sample of the address segments in one panel was selected for the successor check. The first address in each of these segments was used as the address from which the string of eight was determined.

The details of sample selection for the successor check appear below.

2. Systems and Procedures

The successor check was conducted at time of interview for 1,500 selected AHS units. For each of these units the interviewer listed a string of 8 structures[7] in a path of travel bearing to the right from the sample unit. The structures along the route were listed and a sketch drawn to show their location.

The year of construction was determined in order to identify regular structures built before April 1, 1970. These were called successor structures and were used to bound the string. By design, each string was to consist of eight successors and any intervening structures.

The string could cover one or more blocks in urbanized areas or a distance up to 10 miles in rural areas. In general, the path of travel was expected to proceed around the block in which the sample unit was located. In order to preserve probabilities, the string was terminated in the sample unit block when the northwest-most corner was reached. From this point the interviewer would continue an incomplete string, starting at the northwest corner of the next block to the right. This procedure would be continued until the string was completed.

No procedure was required for matching against the census address registers because the operation was not designed to identify census misses. (In address ED's units missed or otherwise not reported in the census are represented through the CEN-SUP

sample.)

Interviewers recorded the number of units in each inscope structure listed. The regional office clerk reviewed the listing sheets and performed various quality checks. Units in inscope structures were assigned for AHS interview. Large multi-unit inscope structures were subsampled.

Consideration had been given to conducting interviews at inscope structures at the time they were identified in the successor check. This had some advantages in relation to cost and time constraints. However, it was felt that this might introduce interviewer bias. If interviews had to be obtained at each inscope structure, interviewers might be less scrupulous about identifying such structures.

A total of 44 inscope successor units were assigned for interview.

VI. RESULTS OF COVERAGE IMPROVEMENT PROCEDURES

The undercoverage bias affected both the total new construction estimates and the estimates of characteristics of the total housing inventory. For each year until 1976, a ratio estimation procedure was employed to adjust the AHS sample estimates of new construction units to independently derived current estimates.[8/] This procedure was used to correct for known deficiencies in the three categories of new construction represented in the sample.[9/] Although the independent estimates were considered the best available, their accuracy had become a matter of growing concern. In addition, the ratio estimation procedure may have had no effect on the bias in housing characteristics due to the undercoverage of certain types of units. The coverage improvement procedures addressed both of these issues. If the procedures could correct frame deficiencies so that all housing units had a known non-zero probability of selection in the survey, this would eliminate the bias and, in addition, valid unbiased estimates of total could be derived from the survey data itself. Another possible option relates to the third stage ratio estimation procedure. It is designed to adjust the AHS total inventory estimates to current independent housing estimates. These latter are derived from the CPS and the HVS. These two surveys have the same frame deficiencies as the AHS. Better estimates of the total housing inventory might be obtained by correcting the frame deficiencies in the CPS and HVS and then retaining the third-stage ratio estimation procedure.

In order to evaluate these options it is first necessary to examine the results of the coverage improvement procedures in terms of their effect on the undercoverage bias in the AHS sample.

The four coverage improvement procedures yielded a total of 1,667 unweighted units, of which 1,538 would be weighted to represent omissions in the housing inventory. The distribution by source and an analysis of their contribution to the sample appears below.

A. Permit Lag Sample Results

There were 1,386 units selected for the permit lag sample, representing 598,000 units which had no other chance for selection in AHS. This amounts to 0.88 percent of the total 1970 housing inventory and is all new construction.

The basic weight assigned to the permit lag sample units, during the AHS weighting procedure, was equal to the inverse of the probability of selecting a sample unit for the AHS permit lag survey.

The weight assigned to each of the sample units in the i^{th} cluster of the j^{th} permit was euqal to $\frac{1316.}{N_{ij}}$.

Originally, the AHS permit lag sample design was to produce a self-weighting sample with each sample unit having a weight of 1316. However, during the AHS permit lag sampling operation the clusters were assigned the measure of size $\frac{N_j}{K}$ rather than $\frac{N_j}{K_j N_{ij}}$, which produced the actual non-self-weighting sample.

The permit lag coverage improvement procedure was probably the most successful in terms of eliminating the undercoverage bias associated with AHS. Since the permit lag sampling frame was based on a representative national 1-in-100 sample of all permits authorized before 1970, it should also be a representative sample and produce approximately unbiased estimates of any subset of these permits. Thus, one would expect that the permit lag sampling frame was a representative sample of units for which construction was authorized before 1970 but was completed after April 1970, and that a sample selected from this frame would produce unbiased estimates of characteristics of such units. The problem was that two possible sources of bias were introduced into the sampling operation. One resulted from eliminating units in non-AHS PSU's from the permit lag sampling frame and the other was the result of noninterviews in selected clusters. However, any bias in the sample estimates from these sources are likely to be quite small. In the first instance the number of units represented is about 16,000 and the weights on the remaining units in the permit lag sample frame were increased to represent these units. The second source of bias resulted from the fact that 21 of the 479 clusters selected for the AHS permit lag sample could not be visited because the corresponding SOC questionnaire, which contained the address, could not be located. Once again, the weights for the sample units that were visited were increases to reflect these 21 clusters. It is fairly safe to assume that these were approximately random misses and thus most of the bias associated with this problem was eliminated by the adjustment.

Although estimates from the permit lag sample are subject to sampling error, the magnitude of the sampling variability is probably lower than it would have been if these units were represented in the original AHS sampling operations. The decrease in variance was due to the larger-than-planned size of the AHS permit lag sample.

This gain was offset slightly by the fact that the permit lag sample frame was based on the 122 PSU design, which is therefore subject to more between-

PSU variance than the AHS 461 PSU design. Also, since the overlap between the permit lag universe and the address segment universe for multi-unit structures[10] was resolved at the sample unit level rather than the universe level, there may have been an increase in the variances associated with the sample estimates from the permit lag universe.

B. Woodall and Windshield Sample Results

A total of 237 mobile home sites located in parks were selected for sample from these two sources. This represented 342,000 mobile home units that had no other chance for selection in AHS. Approximately 50 percent of these mobile home units were in parks that existed in 1970 but were not reported in the census; the remainder were in parks created since the census.

The basic weights assigned to each Woodall or Windshield sample unit during the AHS weighting procedure was equal to 1,366. Thus both the Woodall and the Windshield samples were self-weighting sample designs.

The combination of the Windshield and Woodall coverage improvement procedures was successful in terms of eliminating the mobile home undercoverage bias in AHS. This was due, in part, to the supplemental effect of the Windshield Sample. The Woodall sampling frame consisted of what was purported to be a complete listing of new mobile home parks that were created between April 1970 and December 1974. Thus, the sample selected from this listing should be a representative sample and produce approximately unbiased estmates of that universe. However, there was evidence of undercoverage in the Woodall frame which was improved by the Windshield procedure. This latter procedure was able to represent not only mobile home parks built after 1974 and mobile home parks missed by the census but also mobile home parks that should have been on the Woodall list but were not. Twenty-three of these parks were picked up in the Windshield sample. Thus, the Windshield procedure attempted to eliminate the undercoverage bias in the Woodall procedure due to the deficiencies in the Woodall sampling frame.

As a result, the major source of bias associated with the Woodall sample estimates, i.e., an incomplete universe, may have been eliminated, depending on the bias associated with the Windshield procedures. Although the Woodall sample was selected at the same rate as regular AHS, the Woodall estimates are probably subject to more sampling error than if they had been sampled with regular AHS since the Woodall sample was confined to the 266-PSU design and therefore is subject to more between-PSU variance.

One source of nonsampling error associated with the Windshield estimates is the completeness of the canvassing and of the matching operations. Additionally, since tracts were used as the areas to be canvassed for Windshield, this procedure only represents missing mobile homes in address ED's in tracts. The magnitude of this bias, obviously, depends on the proportion of missing mobile homes in non-tracted address ED's, which are approximately 9 percent of all address ED's. This bias would also impact on the effectiveness of the Windshield in terms of eliminating the undercoverage bias in the Woodall sampling frame. Even though the Windshield sample units were selected at the same rate and in the same PSU sample design as regular AHS, the resultant estimates are probably subject to more sampling error than if these units had been sampled with regular AHS. One source of this additional variance is the fact that the existence of parks in area ED's within tracts could not be corrected for until after canvassing, rather than before selecting the sample of tracts. Thus, the measure of size used in the selection of tracts included the effect of mobile homes in area ED's. The other major source of additional variance is the degree of effectiveness of the measure of size, assigned to the tracts during the selection of tracts, with respect to estimating missing mobile homes.

Missing parks were found in tracts with measures of size ranging from as low as 48 to as high as 2,435, whereas some tracts with measures of size as high did not contain missing parks. Thus, comparing the measures of size for tracts with and without missing parks does not uncover any obvious patterns. Nonetheless, the estimated correlation coefficient, based on these 150 tracts between the measure of size for a tract and the number of sites in missed parks found in the tract, is .67. Thus, based on the magnitude of this estimated correlation coefficient, it appears that the measure used in the selection of tracts was fairly effective in terms of the characteristic of interest.

C. Successor Check

The three types of inscope units discovered through the successor check produced a total of 44 sample units distributed as follows:

28 units - mobile homes at large, of which 24 represented omissions in the housing inventory

11 units - houses moved into the sample area

5 units - converted from nonresidential use

These represented roughly 124,000 units, which have unusual characteristics that were not adequately reflected in the original sample. (The total weighted figure would be 140,000 but 16,000 would not be considered part of the housing inventory.) The basic weight assigned to each successor check sample unit, during the AHS weighting procedure, was equal to the inverse of the probability of selecting the unit. The probability of selecting a sample unit was equal to the probability of selecting a successor check structure. As was mentioned before, the successor check sample design involved the listing of a string of addresses starting from the first address (referred to as "the sample address") in half of the address segments in one panel of AHS (panel 3). The string included exactly eight census addresses which had a prior chance of being selected for AHS (referred to as "the successor addresses") and any intervening new construction, mobile home parks, other types of special places, and inscope structures.

All units in an inscope structure (referred to as "the successor check sample units") were inter-

viewed for AHS unless there was an excess number of successor check sample units in an inscope structure or the string. In that case, a subsample of the successor check sample units was selected for interview.

From this design, every successor address had a chance to be a sample address and vice versa. As a result of listing eight successor addresses in each string, any inscope structure could have been brought into the sample because of one of eight possible AHS sample addresses. If the eight preceeding sample addresses (or equivalently, successor addresses) for an inscope structure are denoted by $a_1, a_2, \ldots, a_8$ and the probability that sample address a_k was selected for the successor check is denoted by $P_r[a_k]$, then the probability that an inscope structure came into sample is $\sum_{k=1}^{8} P_r[a_k]$. However, information was obtained such that $P_r[a_k]$ could be calculated for only those successor addresses that preceeded the inscope address in the string. Therefore, $\sum_{k=1}^{8} P_r[a_k]$ could not be calculated from the information available. Nonetheless, the conditional probability of inclusion of the inscope address, given that the sample address is a_k, did provide an unbiased weight.

This conditional probability of inclusion is $8\,P_r[a_k]$ and was estimated as follows:

Let q = the number of address segments with some or all of their units in the sample address a_k.

m_i = the number of units in the i^{th} address segment at the sample address.

n_i = the total number of units in the i^{th} address segment.

Therefore:

$$8\,P_r[a_k] = 8 \sum_{i=1}^{q} P_r\left[\begin{matrix}\text{Panel 3 was selected}\\ \text{for successor check}\end{matrix}\right] \times P_r\left[\begin{matrix}\text{The half-panel}\\ \text{was selected for}\\ \text{successor check}\end{matrix}\right] \times P_r\left[i^{th}\ \text{segment was selected for AHS}\right] \times P_r\left[\text{sample unit falls in } a_k\right]$$

$$= 8\left\{\sum_{i=1}^{q} \frac{1}{6} \times \frac{1}{2} \times \frac{2}{1366} \times \frac{m_i}{n_j}\right\}$$

$$= \frac{4}{3(1366)} \sum_{i=1}^{q} \frac{m_i}{n_i}$$

Thus, the basic weight for the units in an AHS successor check inscope structure was equal to:

$$\frac{3}{4}\left(\frac{1}{\sum_{i=1}^{q} \frac{m_i}{n_i}}\right) 1366$$

Thus, the successor check basic weight was

$$\frac{3}{4}\left(\frac{1}{\sum_{i=1}^{q} \frac{m_i}{n_i}}\right)$$ times as large as the regular base weight for AHS sample units from address segments. The most common successor check basic weight was 3.0 times the regular AHS base weight.

There is evidence that the successor check coverage improvement was not very effective in terms of eliminating the undercoverage bias in AHS for the types of units involved. As was mentioned before, the 1970 CINCH Survey estimated that additions from other sources, which are comparable to the units represented by successor check, between April 1960 and December 1970 amounted to 1 percent of the 1960 census inventory. Extrapolating this rate to the time period April 1970 - October 1976 indicates that additions from other sources in this time period should be about .6 percent of the 1970 census inventory or 400,000 units. (This is probably an underestimate of the actual rate for a 6½ year period because the CINCH estimate which covers a 10 3/4 year, would not represent units added by other sources after 1960 that were removed from the inventory by December 1970.) Since the successor check was designed to represent these types of units in address segments only, this number should be adjusted by the percent of the old construction sample represented by address segments (75 percent). This produces a figure of about 300,000 units which, even though it is probably an underestimate of the actual figure, is substantially larger than the estimate from the AHS successor check sample. Most of this difference was probably due to the successor check's coverage of the units converted from nonresidential to residential use. For these types of units, the successor check sample produced an estimate of about 16,000 units, which seems extremely low for a 6½ year period.

In evaluating the results of the successor check two important matters of resource must be considered. First, better estimates would have been obtained by selecting a sample from a representative list of these units, such as was done for the permit lag and the Woodall sample. The problem was that no such list existed or could be compiled with available resources. A second approach would have been to use the CINCH (area block listing) method. However, this would have been a costly operation and would have taken more time to finalize than was available for AHS. As a result it seemed best to develop a procedure that could be integrated into the basic AHS sample. The successor check was a reasonable choice for identifying mobile homes at large and even homes moved in. However, some type of stratification, or a very large sample might be required to provide an adequate sample and control excessive variability for nonresidential conversions. For example, to alleviate the deficiency, the

successor check sample could have included a disproportionate number of AHS sample units in nonresidential areas. Alternately, a block sample approach could have been used for nonresidential conversions, in which a sample of blocks within a sample of tracts, which were highly nonresidential in 1970, could have been canvassed to identify such units. However, either of these methods would have added to the costs and funding was a serious problem. In any case, further consideration will have to be given to this matter before the successor check can be introduced into other surveys.

VII. SUMMARY AND CONCLUSIONS

A number of known deficiencies existed in the sampling frames for the AHS National sample. These resulted in under-representation in the survey of mobile homes, new construction, housing converted from nonresidential structures and houses that were moved from their original sites. The total estimate of this undercoverage was about 959,000 units. Although the survey data were adjusted to compensate for these omissions, biases may still have existed in the characteristics of the housing inventory. In addition, the independent estimates employed in the estimation procedures were not entirely satisfactory, especially for new construction mobile home estimates. Therefore, supplementary coverage procedures were introduced into the 1976 survey to provide more adequate coverage in the survey itself. The results of these procedures in terms of the undercoverage bias were the subject of this paper.

Since the coverage improvement procedures virtually eliminated the undercoverage bias for new construction units, it was decided to eliminate the second-stage ratio estimation procedure for most categories in the 1976 AHS estimation process. However, undercoverage bias in the AHS sample still exists for units converted from nonresidential use since 1970 in address ED's and for units in area ED's, so some concern still existed about the estimation of the total inventory. Therefore, it was decided to continue using the third-stage ratio estimation procedure for the 1976 AHS, even though it was felt that the independent estimates were overstated. This is a conservative approach and is subject to change in later years, as more experience is gained from the use of the coverage improvement procedures.

1/ Some additional sample units (representing 121,000 units) that were picked up by the coverage improvement procedures are not included in this figure, although they could become part of the housing inventory. These include vacant mobile homes and unoccupied sites in mobile home parks that may be occupied in the next AHS interview period and therefore included in the housing inventory at that time.

2/ For example, vacant mobile homes or sites in parks were not recorded in the census but are included in AHS since the units might be occupied when the annual survey is conducted.

3/ Inadvertently, during the actual sampling operation, the $N_{ij's}$ were not divided out. This meant that the probability of selection for sample units was N_{ij} times as large as had been intended to produce a self-weighting sample. This produced a larger-than-expected sample for this coverage improvement procedure which necessitated a special adjustment in the estimation procedure for these units.

4/ Details of this process can be found in the 1970 redesign documentation, which is mostly internal Census Bureau memoranda. They will also appear in the Bureau's Technical Paper No. 7 which is currently being revised.

5/ This refers to Census tracts, which are geographic areas containing two or more ED's.

6/ Tracts in which no mobile homes were reported in the 1970 census were assigned a measure of size of 1 to insure that these tracts had a chance of selection.

7/ The term structure, as used here, includes mobile homes at large on permanent foundations if occupied by persons with no usual residence elsewhere, as well as regular residential structures.

8/ For more detail on the estimation procedure see Current Housing Reports Annual Housing Survey: 1973 United States and Regions, Part A, Series H-150-73A, Appendix B, "Sources and Reliability of the Estimates," pp. app. 32-3.

9/ This included categories for conventional new construction units and for new mobile home placements.

10/ The situation is described in section V.A.2. of this paper.

NOTE: This paper is an abstraction which omits details of the sample design and estimation procedures, related research, and references that appear in the original paper.

SOME LESSONS LEARNED FROM SSA EXPERIENCE IN CONTRACTING FOR SURVEYS*

Thomas B. Jabine, Social Security Administration
Nathaniel M. Pigman, Jr., Health Care Financing Administration

ABSTRACT

Much of the Social Security Administration's program-oriented research and program evaluation is carried out through surveys conducted for SSA by the Bureau of the Census or by private contractors. The Statistical Methodology Group of SSA's Office of Research and Statistics conducted an in-house study of survey management procedures, giving special attention to the development of survey design specifications through interaction of the sponsoring agency and the survey organizations. The study procedures are described and some findings are given. A suggested checklist for use in the preparation of technical scope of work statements for survey RFP's is presented and discussed.

INTRODUCTION

Background - The number of surveys sponsored by Federal agencies has increased rapidly in recent years. With a few exceptions, the survey data are not collected and processed by the sponsoring agency. The work is either done by a private survey organization under contract, normally executed as the result of competitive bidding, or by another Federal agency under a reimbursable agreement. The Federal agency doing the greatest amount of reimbursable survey work is, of course, the Bureau of the Census.

During the past 2 or 3 years, several organizations have exhibited serious concerns about the quality of Federally sponsored surveys. The Subsection on Survey Research Methods of the American Statistical Association, with funding from the National Science Foundation, has recently completed a feasibility study for a project on the assessment of survey practices (Bailar and Lanphier, 1977). The findings were disturbing - for the 26 Federally sponsored surveys included in the study (a purposive sample), it was found that 10 failed to meet their objectives, 11 did not use probability sampling throughout, 4 had designs rated as poor by the investigators, and 15 either had response rates of less than 75 percent or their response rates could not be determined. On the basis of the feasibility study findings, ASA has applied to NSF for funding for a larger study, to be based on a probability sample of all surveys conducted during the reference period selected.

Other organizations which have recently concerned themselves with the quality of Federal surveys include the National Center for Health Services Research, the Joint Ad Hoc Committee on Government Statistics (1976), the Federal Paperwork Commission (1977), and the newly-formed Council for Applied Social Research, which has established an annual award for the "Best RFP of the Year".

*Opinions expressed are those of the authors and do not necessarily represent the positions or policies of their respective agencies.

Surveys sponsored by the Social Security Administration - Almost from the beginning of social security, surveys have been an important research tool for the Social Security Administration. Surveys of beneficiary populations are used to study issues such as adequacy of benefits, relation of benefits to income from all sources, and comparative program effects for population subgroups characterized by age, sex, race/ethnicity, education and other demographic and social variables. Surveys of potential beneficiaries (target populations) are used to determine participation rates for different groups, knowledge of programs and reasons for applying or not applying for benefits. As new benefit programs, such as disability, Medicare, and supplemental security income have been added to the original retirement and survivors program, surveys have been used to provide information about the new beneficiary and target populations.

The conduct of these surveys has passed through 3 stages. Initially, all surveys were conducted "in-house," i.e., by SSA district office personnel, according to specifications developed by the research staff in central headquarters. Starting with the Survey of the Aged in 1963, the Census Bureau has conducted several major national surveys for SSA on a reimbursable basis. Finally, since 1970, as the number and variety of surveys has increased, many of the surveys have been conducted under contract by private nonprofit and commercial survey research organizations.

Most, although not all SSA surveys are planned and carried out under the direction of the Office of Research and Statistics (ORS). Primary responsibility for these surveys rests with the program divisions of ORS - the Divisions of Retirement and Survivors Studies, Disability Studies, Supplemental Security Studies and, until recently, Health Insurance Studies.[1/] Typically, the appropriate division director takes overall responsibility for a survey and under his general direction a member of his staff, usually a social science research analyst, is designated as the project manager, and, for contract surveys, as the project technical officer. Each of the program divisions has one or more mathematical statisticians and they are generally called on to assist in various phases, such as survey design, evaluation of technical proposals, and analysis of results.

Direct responsibility for the procurement process for contract surveys rests with the Division of Contracting and Procurement in the Office of Management and Administration. Within ORS, the Office of Research Grants and Contracts provides assistance to ORS divisions in their contracting activities and takes direct technical responsibility for selected projects. HCFA has an individual with similar functions on the immediate staff of the Associate Administrator for Policy, Planning and Research.

A Study of Contract and Reimbursable Surveys - Early in 1976, one of the authors was asked to review and comment on the sample designs included in technical proposals submitted in response to an RFP for a new ORS survey. As a firm believer in the use of probability sampling, he was disturbed to find that some of the proposals did not call for probability sampling at all stages of the design, and he recommended that these proposals be disqualified. However, it turned out that this could not be done because the RFP had not specifically called for probability sampling. At best, these proposals could be given low scores on the relevant selection factors; however, they would not be disqualified from contention on this basis.

This experience led to a recommendation that RFP's for ORS surveys should routinely include a standard clause calling for the use of probability sampling. Prior to its eventual adoption, the proposed standard clause (see Exhibit A) was submitted to the ORS Statistical Methodology Group (SMG) for review. The SMG is an informal group of mathematical statisticians from the various divisions of ORS who meet periodically to discuss applications of statistical methodology in their work and to share experiences and problems. From time to time, ad hoc groups are formed from the SMG to address problems of general interest.

In the SMG's discussion of the proposed standard clause on probability sampling, it was pointed out that there might be other ways in which the quality of SSA contract and reimbursable surveys could be improved. There being general agreement on this point, a working group was established to undertake a study of SSA contract surveys. The object of the study was to review and evaluate SSA procedures for contracting with outside organizations to conduct statistical surveys and to identify those provisions of RFP's and contracts which are instrumental in specifying the quality of the survey design and execution, and to see to what extent such provisions have been fully complied with. It was expected that the study would serve as a basis for developing improvements in our survey contracting process in one or more of the following ways:

- By providing guidelines for preparing technical statements of work for inclusion in survey RFP's, covering such factors as specifications designed to insure high completion rates, required use of probability sampling, calculation of sampling errors, etc.
- By providing appropriate training and technical assistance to staff members preparing such technical statements of work to be performed.
- By suggesting improvements in the contractor selection process.

In this paper, we describe the design of the study and present some preliminary findings and some recommendations for improvement of survey management procedures based on these findings.

THE CONTRACT STUDY: DESIGN AND METHODOLOGY

Defining the Population - For the investigation it was deemed feasible to study all SSA contracts and reimbursable agreements involving surveys for which RFP's were issued and contracts executed for the 1975 and 1976 fiscal years (including the transition quarter). Frame problems were encountered in that some contracts on the original list did not actually involve statistical surveys and others represented follow-up surveys in which the original specifications had been prepared earlier. In one case the contract was essentially an "off-the-shelf" procurement where the survey design had not been specifically developed for SSA but was in place at the time of the contract. In general, the principle evolved that any situation which might provide insight into the area under study was included. From an original list of 19 contracts and reimbursable agreements, 13 were deemed to include surveys suitable for analysis and inclusion in the study.

Development of Survey Instrument - The development of the survey instrument was largely a heuristic process based on a review of several of the contract documentation sets. From these emerged a three-part data collection instrument. The first part was a cover sheet identifying the project, categorizing it as a contract or reimbursable agreement and, in addition, covering survey characteristics such as coverage, sample size and data collection procedures. These data were set up in a format convenient for the abstracting process.

The second part of the instrument was a narrative questionnaire going into considerable detail with respect to the coverage, sample design, frame, sample size, response rates, and collection procedures. A third part of the instrument consisted of a listing of source documents keyed to relevant portions of the narrative.

The instruments required some revision as the study progressed but remained substantially unaltered in content. A copy of the survey instrument may be obtained by writing either author.

Data Collection Procedures - The personnel available for the study, with one exception, participated in the study on a part-time basis--doing as much as other duties permitted. The chief manpower pool for the preparation of the narrative section were members of the Statistical Group described earlier. The principal sources of data were the files provided by the contracting management units and in some cases files provided by analysts who had been actively engaged in the development of the survey. A two-stage process was employed which consisted first of the collection of the source documents and preparation of the cover sheet (Part I of the Contract Study Questionnaire). The chief categories of source documents were planning memoranda, requests for proposal, contract documents, including the technical proposal of the successful offeror, costs estimates, supporting statements to requests for

OMB clearance, and interviewer and training manuals.

From these a full-time analyst prepared the cover sheet and, with the supporting documents, prepared a folder for each of the study contracts.

These folders were then distributed to members of the Statistical Methodology Group who undertook the preparation of the narrative and source document portions of the Contract Study Questionnaire. This activity in some cases involved going beyond the prepared record to ancillary files and discussion with survey analysts. After completion of the study questionnaire the results were distributed for comment to the project officers involved in the survey under review.

FINDINGS

The Study Population - The study population consisted of all ORS statistical surveys for which contracts or reimbursable agreements were executed during 1975 and 1976 fiscal years. The distribution of characteristics of the surveys included in the study is shown in Table 1.

Table 1 - ORS Survey Profiles: Number of Surveys with Designated Characteristics

Type of Agreement	**Status**
10 - Contract	2 - Complete
3 - Reimbursable	10 - Incomplete
Type of Contract	1 - Ongoing
8 - Fixed Price	**Reporting Unit***
4 - Cost Reimbursement	7 - Individual
1 - Cost Sharing	3 - Hospital
Type of Bidding*	5 - Other
4 - Sole Source	**Principal Collection Method***
8 - Competitive	6 - Telephone
Coverage	4 - Mail
11 - National	10 - Face to Face
2 - Other	**Pilot Study**
	4 - Yes
	9 - None or Not Applicable

The sample sizes for these surveys ranged from about 1,000 to 20,000, and the out-of-pocket costs (i.e. for contracts or reimbursable agreements) from a low of about $22,000 to a high of over $3,000,000. To give an idea of what these extremes represent, the $22,000 figure was for a mail survey with telephone followups addressed to utilization review officials in a sample of about 1,000 hospitals. The response rate was slightly under 50 percent. This survey was done as a small part of an evaluation of concurrent utilization review procedures. At the other end of the scale, the $3,000,000 + figure was for the Survey of Low Income Aged and Disabled, a survey in which two personal interviews were conducted, about one year apart, for a sample of about 20,000 persons receiving or potentially eligible for benefits under the Supplemental Security Income program.

Findings for Contract Surveys: Basic Survey Objectives - As described further in the section on "Application of the Study Findings", the main product resulting so far from this study has been a checklist for use in preparing technical scope of work statements for survey RFP's. In that checklist (see Exhibit B), we have listed a "minimum set" of survey objectives which must be provided by the sponsor as a basis for designing any survey. These are (a) definition of the survey population, (b) kinds of information to be collected, (c) use of probability sampling, (d) level of sampling error (reliability) desired, and (e) target response rate. The major findings from the study relate to these 5 items. For each item, we have asked the following questions:

1. Was a formal specification adopted?
2. At what stage in the survey process did the specification first appear, i.e., was it in an RFP, in a technical proposal, a contract amendment, the OMB clearance submittal, etc.?
3. Was the specification adequate?
4. Was it carried out?

Because several of the surveys studied are still underway, we were not always able to answer the last question.

A general finding about the specification of survey objectives was that there was a striking difference between those cases where the survey was the primary purpose of the contract and those where the survey was a secondary or minor part of a contract for an evaluation study. In the latter case, the specifications were much less likely to be clearly documented, and the overall quality of results, to the extent it was ascertainable, was in general less satisfactory.

(a) Definition of the survey population - The target population was judged to be well-defined in nearly all cases. Typically, this was covered in the scope-of-work section of the RFP. Two issues emerged:

(1) For national surveys (which most of ours were) the final result was frequently a probability sample in which members of the target population in the States of Hawaii and/or Alaska were given no chance of selection. In Census Bureau terms, the study was limited to the population living in the conterminous United States.

Obviously, this was done to keep costs down. However, residents of these 2 States might have a legitimate complaint if they are routinely excluded from most surveys.2/ Also, it suggests that some care should be taken in evaluating the costs of alternative proposals where the offerors have established national samples of primary units in which they propose to conduct the survey. The offeror who has excluded Hawaii and Alaska from the universe is offering a different product and one which intrinsically has a lower cost per interview.

(2) There were some ambiguities in defining the relationships between individual members of the target population, ultimate sampling units and reporting units. Usually but not always there is a one-to-one correspondence among all 3 types of

*Some surveys combined more than one classification

units. In some surveys there may be more than one type of reporting unit, e.g., individuals receiving SSI benefits and recipient units, such as a husband and wife receiving SSI benefits. Surveys related to income maintenance programs may deal with many kinds of units, including individuals, beneficiary units, families and households. If the reporting unit contains more than one person, it may be necessary to interview more than one person to collect the desired information. Therefore, in any discussion of sample size and/or number of interviews it is necessary to be precise about the kinds of units being discussed. Also, if the sampling frame consists of individuals, more than one of whom may be members of the same reporting unit, the multiple probabilities of selection for some reporting units must be taken into account in preparing estimates from the survey.

(b) Kinds of information to be collected - The documentation and adequacy of content specifications was not directly addressed in the study questionnaire; therefore, we do not try to present an overall evaluation. However, there are some relevant comments that can be made:

(1) In several cases, a draft questionnaire was made part either of the RFP, or of the technical proposal presented by the successful offeror. Putting the draft in the RFP gives the offeror a good basis for estimating costs of collecting and processing the data.

Processing costs are significantly increased by the inclusion of open-ended or unstructured questions, so the offeror needs to know whether and how these will be used.

(2) For various reasons, it may be useful to designate one or more key variables, representing the most important results to be obtained from the survey. These key variables may then be used to specify requirements for sampling reliability and also in the process of deciding whether or not an interview or questionnaire may be counted as "complete". Where appropriate, the definitions of these variables should include geographic (national, regional, State, etc.) and time (level or change) dimensions.

Most of the surveys reviewed in this study did not explicitly define key variables. We have seen one instance of an RFP for a periodic survey (issued after the end of the reference period for this study) where failure to specify the relative importance of estimates of level vs. estimates of change led to considerable difficulty in making a comparative evaluation of the survey designs proposed by different offerors.

(c) Use of probability sampling - The standard clause on probability sampling for RFP's (Exhibit A) was developed subsequent to the award of contracts for the surveys included in this study. Nevertheless, the record was almost uniformly good concerning the use of probability sampling in these surveys.[3/] The one clear exception was a case in which we contracted to obtain data on prices of drugs purchased by pharmacies from an ongoing market survey. A careful review of the selection procedures by ORS statisticians after we had been using the data for several months made it clear that some members of the universe had no chance of selection and that it was impossible to determine exact selection probabilities for stores in the sample.

Probability sampling was a specific or implied requirement in the RFP in about half of the contract surveys studied. In other cases, its use was specified or documented at a later stage, e.g., in the successful proposal, in the contract, or in a contract amendment. Several contracts provided for an agency review of the proposed sample selection plan prior to execution. This has proved to be an effective method of avoiding unintentional departures from probability sampling and in some cases has led to more efficient designs.

Probably the most important lesson we have learned about probability sampling in this study and through experience with contract surveys is to be extremely careful when "buying in" to previously selected samples. Before agreeing to the use of a particular sample alleged to be a probability sample, agency representatives should insist on making a critical review of the design specifications and of the actual sample selection worksheets or other relevant materials. If the proposal calls for some modification of an existing sample used by the offeror, plans for such modifications should be fully reviewed. Modifications frequently proposed include expansion or subsampling of an existing sample or use of a set of PSU's designed for an area sample to select a sample from a list of program participants. In the latter case, if the participant list does not carry county codes, appropriate procedures or rules must be developed for associating each unit on the list with a particular county or other geographic unit used to define the PSU's in the area sample. From the point of view of sampling efficiency, if the distribution of program participants is not reasonably well correlated with the measures of size used by the offeror to select his PSU's, a larger sample will be needed to obtain the desired reliability of estimates.

(d) Level of reliability desired - Most of the RFP's for contract surveys took the more or less traditional approach of requiring a specified number of completed interviews. One or two also specified that these interviews be conducted in some minimum number of PSU's. Strangely, in one case the sample size was not specified at all in any of the procurement documents and in another, a rather wide range was given. We have also noted that if the sample size is not clearly specified in the RFP in terms of completed interviews (or alternatively, as the initial sample, with a minimum or target response rate), some offerors will treat it as the initial sample and some as the number of completed interviews.

Clearly, when requirements are given in terms of probability samples of fixed sizes, it is possible for offerors to meet these requirements with sample designs which vary substantially in terms of their expected reliability for estimates of specified population values. If these are the only RFP requirements relevant to reliability, the designs which produce less reliable estimates will, in general, tend to have lower costs, and thus be

at an advantage in the selection process.

This is not a simple problem to solve. Ideally, we might specify the desired reliability for a few key variables. In practice, this may be difficult for many reasons. We may not know enough about components of variance for these variables to set target reliabilities which can be reached within the budget allotted to the project. It may be difficult to persuade the users (in-house) of the data to select key variables and specify target reliabilities. Finally, it may be difficult to decide whether or not proposed designs will meet these targets. Nevertheless, we believe that this approach should be used when feasible.

In a recent survey RFP, we required that the proposed design produce estimates with reliability equivalent to estimates from a simple random sample of a specified size. This may be a useful procedure where most of the significant estimates from the survey will be proportions or percents based on attributes. We are not yet at liberty to discuss the results; however, we can say that the experience has shown that there is a dearth of specific data on design effects for different survey designs and variables.

(e) Target response rate - The study shows this to be the area which was neglected most in the procurement process. None of the RFP's specified a target response rate[4]; insofar as we could determine, only one specified in any detail the required efforts to obtain complete response.

In a majority of cases, an expected response rate or a reasonably complete description of the planned followup effort or both appeared either in the technical proposal (which is incorporated into the contract) or in a contract amendment. In one case, the contractor planned to review response rates for different cells based on respondent characteristics and do telephone followups for cells where response was low. We did not consider this to be a description of an adequate followup effort.

For one survey, we could not find any information about response rates until we reached the OMB clearance submittal. There we found both an expected response rate and a detailed description of planned followups! There may be a moral here for those who contend that the OMB clearance process is a waste of time.

With respect to actual performance, we have only partial information. This is partly because several of the surveys are still underway, but also is due in part to failure to document response outcomes fully.

Where we do have information we find that our surveys of beneficiary and target populations usually achieve a reasonably high response rate, but that the experience with surveys of health care providers, e.g., hospitals and physicians, has been less satisfactory. While this difference may be attributed in part to intrinsic difficulties in surveying the latter group, we believe it also results partly from giving insufficient attention to response problems during the procurement process.

Finally, in connection with followup effort, it is important to remember that costs are directly related to the amount of followup effort. As was the case for reliability, we must, in the procurement process, avoid giving an unfair advantage to the offeror who proposes a minimal or vaguely defined followup effort.

Findings for Contract Surveys: Deliverables- Part B, 1 of the Checklist for RFP's (Exhibit B) lists several possible "deliverables", i.e., concrete work products that are required to be delivered to the agency by the contractor at specified times. In some cases, these items must be approved by the agency before later stages of the survey process can start.

Most of these items were included in the majority of contracts studied. However, there were 4 items - h, i, j and m - which were rarely found in contracts. Significantly, these were all items which provide information about the quality of the survey results. It is almost as if we have been saying to contractors "Give us the data and the analysis on a timely basis, but don't tell us anything about errors in the data." Following is a brief discussion of these 4 items:

1. (Item h) A detailed and accurate accounting of the data collection results for the initial sample. This information is needed in order to

(a) Determine how well the contractor succeeded in meeting target response rates.

(b) Make appropriate adjustments for nonresponse in producing estimates from the survey data.

(c) Advise data users about potential nonresponse errors in the results.

(d) Set reasonable targets for response in later surveys.

2. (Items i and j) Quantitative information on the results of validation and verification in the data collection and processing operations. Most contracts provide for validation of a sample of the interviews conducted and for 100-percent or sample verification of coding and keying operations. However, we seldom ask for or receive information on the findings of these checks. Asking for such data might increase the probability that these checks would be taken seriously, and would provide further information of interest in connection with the analysis of the results.

3. (Item m) Estimates of sampling error. ORS has a policy of presenting sampling errors when results based on samples are published. However, the need to calculate sampling errors sometimes doesn't occur to the survey manager until fairly late, e.g., when the tabulations are completed and it is time to analyze the data and prepare a report. Consequently, we find that the contract seldom provides specifically for the calculation of sampling errors. In some cases this is deliberate, as we plan to do the calculations ourselves; however, even in such cases it is important to insure, through appropriate contract provisions, that the data turned over by the contractor include the information needed to calculate sampling errors based on the sample design actually used.

Findings for Reimbursable Surveys - Three reimbursable surveys were included in our study. A fourth was in scope but we have not yet compiled the relevant information. In all 4 cases the Bureau of the Census was the service agency and was completely responsible for data collection. Responsibility for the selection of samples depended on the frame used. If the frame was a list of SSA program participants, SSA selected the sample; if the frame was a Census or the Current Population Survey, the Census Bureau selected the sample according to agreed-on specifications. Responsibility for data-processing varied all the way from complete processing of questionnaires through the tabulation stage by Census to just the reverse. The confidentiality requirements for Census and Current Population Survey data are a factor in determining these arrangements. One of the 3 surveys included in the study is a continuing survey; the other 2, and the one not included are all longitudinal surveys, i.e., they involved 2 or more interviews with the same respondents.

With respect to the basic survey objectives discussed under the findings for contract surveys, we can make the following observations:

1. The survey population and kinds of information to be collected are usually fully and clearly specified, although not necessarily in a formal way.

2. Probability sampling is always used; both Census and SSA/ORS rely almost exclusively on probability sampling in their survey work.

3. Sample size is usually specified in terms of number of persons or households in the initial sample and number of PSU's. Since variance data on design effects are fairly readily available for Census PSU designs, this is equivalent to specifying reliability.

4. No target response rate is specified, and as far as we could determine, interviewer instruction manuals are not specific about the followup efforts, although general instructions for planning callbacks are included. Nevertheless, response rates, where known, are generally high. Response rates are normally reported in detail for the main survey, but it is sometimes difficult to determine the effects of nonresponse in preliminary screening operations or in the collection of data for the sampling frame (Census of Population or Current Population Survey). Often the combined effects of undercoverage in the frame and nonresponse in all phases leading up to and including the main survey are greater than is generally realized or reported.

An interagency reimbursable agreement is executed for each fiscal year in which work is carried out by the service agency. The description of the work to be done is usually much shorter and less detailed than a contract for a survey. Typically, it does not incoude a detailed time schedule for the work to be done and for "deliverables." Other documentation varies from one survey to another, depending on arrangements worked out between the staff of the 2 agencies who are responsible for the project. Interagency memoranda or letters are commonly used to transmit and react to more detailed specifications. For some of these surveys, we found it difficult, after the fact, to obtain information about all aspects of the survey design.

There are no easy answers in making a choice between the contract and reimbursable routes for a particular survey. It is probably fair to say that the sponsor has at least the potential for more direct control over and ability to monitor the survey operations with a contractor than he does where the work is done by Census. The contractor has a firm legal obligation to perform; whereas the Census Bureau must give priorities to the requirements of its own census and survey operations.

On the other hand, Census offers important advantages, including an experienced and well-supervised data collection staff, access to efficient sampling frames for surveys whose target populations are relatively small and scattered among the general population, and technical resources matched by only a few private survey organizations.

APPLICATION OF THE STUDY FINDINGS

The most important product of this study so far is the Checklist for RFP's for Contract Surveys (Exhibit B). We still regard the Checklist as preliminary and we hope, by presenting it to several reviewers and audiences, to receive numerous suggestions for improvement. Evaluation is needed from both agency sponsors and contractors, and from both survey technicians and analysts, and specialists in contracting procedures.

To make the Checklist more or less self-contained, we have included an introduction describing the general structure of an RFP for a survey. More detailed information explaining contracting procedures to the layman are available from several sources (cf. U.S. Department of Health, Education, and Welfare 1971, 1975).

The Checklist is already being used informally in connection with some RFP's for new surveys. If it stands up after review and informal testing, we expect to recommend, for our respective agencies:

1. That the Checklist be distributed to all current and potential survey managers and project technical officers, and that seminars be conducted for staff members to explain, illustrate and discuss its use.

2. That every survey RFP be reviewed, prior to issuance, by a qualified user of the Checklist.

We have chosen to concentrate on this phase of survey management because we believe that there is no acceptable alternative to building in quality at the beginning of a survey.

While we believe that use of the Checklist will lead to some improvements, it will certainly not solve all the problems associated with survey procurement. Some of these are discussed in the next section.

SOME UNRESOLVED PROBLEMS

Based on our findings in this study and on recent direct experience with the procurement process, we have identified two aspects of survey procurement which we believe require special attention. The first of these - the establishment of response rate requirements - is peculiar to surveys. The second - the selection process - is, of course, much broader in scope.

Response Rate - The establishment of response rate requirements on close analysis becomes a tangled thicket. Interconnected are problems involving potential harassment of nonrespondents, the burden imposed on respondents, and measurement of the incremental benefits derived in terms of total survey error.

Contractual approaches to securing required response rates are varied but not of a nature to totally guarantee results. For example, in the context of the fixed price contract several approaches are possible. For a given price, a specified initial sample size and level of response may be required. Should the company fail to meet these requirements there is no payment.

Not a very satisfactory situation! Another approach would be to establish a variable payment rate tied to the level of response obtained. This incentive approach leaves the financial commitment uncertain but may be more equitable. However, unless coupled with a minimum response requirement, it also leaves the ultimate response rate highly uncertain. Another approach would be to establish a minimum level of accomplishment, and to impose penalties in terms of reduced payment for failure to reach this level.

The most extreme contractual approach to the level of response problem is the employment of the cost plus fixed fee contract. This may be coupled with incentives also, but the chief feature is the commitment to cover all costs associated with the effort. Bluntly, the Government pays the costs or the company stops work.

An indirect approach to response level is the provision in the contract of specific procedures and effort to be exerted in followup of nonresponse. This would include the number of followup visits, telephone calls, or communications required to meet contract requirements. These could be included under the various types of contracts discussed above. Unless the follow-up procedures are rigorously specified, their effectiveness may vary substantially depending on how they are interpreted.

In some situations nonresponse becomes a specific element in the sample design with provision made for double sampling with followups.

The general principle of accountability can be rendered explicit in the RFP as explained earlier in connection with the findings about "deliverables" in contract surveys. Specifically, the contract should require a full accounting of the data collection results obtained for the initial sample, as described in B, 1, h of the Checklist (Exhibit B). Inclusion of this requirement may be expected to provide an incentive for better response results.

We have not had enough experience with different methods of specifying response targets to reach any general conclusions. Further experimentation with alternative approaches is needed.

The Selection Process - Some of the problems considered above as well as problems associated with the evaluation phase of competitive proposals may conceivably be dealt with by a restatement of the entire process. This approach is offered as a beginning, tentatively, and hopefully, recognizing that it may be substantially at variance with existing procurement regulations and policies.

Under the present procedures for negotiated contracts, both technical quality and price enter into the selection process, and their respective weights in the final decision are not always clear.

Our proposal is that the price of the contract be fixed and that the selection be made solely on the basis of technical quality. Thus, offerors would be informed in precise terms of the objectives and the exact budget for the survey and asked to submit technical proposals which, in their opinion, would minimize total survey error for designated key variables.

RFP's, under this system, would not be very different. The scope of work statement would still describe, in fairly precise terms, the target population, the kinds of data required, the time schedule, and specific items to be delivered to the agency. Instructions for technical proposals would specify items to be described by the offeror, including sample design, data collection procedures, data processing and analysis procedures, quality control techniques to be applied, relevant experience of the organization and identification and experience of staff to be assigned to the project.

One important difference would be that the sample size, sampling variability and target response rate would not be included in the scope of work, nor would the use of specified data collection and processing procedures. Each offeror would, however, be expected to cover these items in his technical proposal and to justify his proposed design, as well as to present the usual schedules of work and man-hour allocations by function.

The technical evaluation would become the key to the selection process. Evaluation factors would not differ greatly from those currently in use, but they should cover all possible sources of error in the data, with weights assigned in proportion to the expected importance of each source of error. Specific factors covering sampling error (a function of the proposed sample design) and expected nonresponse error (a function of the proposed data collection procedures) should be included.

Preliminary ratings would be assigned to the proposals submitted and, by a process similar to that now in use (or possibly just by using a numerical cutoff), clearly inadequate proposals would be eliminated as technically not acceptable.

Where necessary the remaining offerors would be contacted, but solely for the purpose of clarification, not for modification of their proposals.

Final ratings would be assigned and the proposal with the best rating would be selected.

The above is an over-simplified outline of a complex process, and undoubtedly it would require some changes and additions in order to function well. A key consideration is the qualifications of the technical evaluation panel. Members should be well-versed in both the theory and practice of statistical surveys. Not all agencies have this kind of expertise in-house; if not, it should be sought from outside.

We cannot pretend that this process would always buy the best (minimum total error) product for the agency. Factors contributing to errors in surveys are many and their individual and joint effects on total error are not fully predictable. However, we believe that selection based on technical merit rather than price would, over time, upgrade the quality of contract surveys (which presently is not all it should be) and would simplify the contracting process in important ways.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the important contribution of the following SSA and HCFA staff members who collected and compiled the information for the 13 surveys included in the study: Erma Barron, Bennie Clemmer, Miles Davis, Tom Herzog, Peter Klein, Bob Mitchel, Pat Moore, and Manny Schwartz. We also want to thank Henry Brehm of SSA and Sid Galloway of HCFA for assistance in identifying the surveys to be included in the study, and the many staff members of both agencies who assisted us in locating the necessary documents and records, and patiently answered difficult questions about the projects for which they were responsible. Finally our thanks to Wayne Finegar, who was technical officer for the first ORS contract survey, for his valuable comments on an earlier version of the Checklist.

FOOTNOTES

1/ Under the 1977 reorganization of DHEW, the Medicare program was transferred to a new agency - the Health Care Financing Administration. At the same time, the AFDC program was transferred into SSA, so in all probability ORS will be conducting surveys of its beneficiaries and target population.

2/ Alternatively, those who consider most survey research to be an unwarranted invasion of privacy might be delighted!

3/ In this section we are concerned only with the intent to use probability sampling. Some of the surveys had low response rates raising questions as to whether the data actually obtained could be characterized as probability samples.

4/ Some later RFP's have included response targets.

REFERENCES

Bailar, B. A. and Lanphier, C. M.

1977 Development of Survey Methods to Assess Survey Practices: A Report of the American Statistical Association Pilot Project on the Assessment of Survey Practices and Data Quality in Surveys of Human Populations. National Science Foundation Grant No. SOC 74-22902, Washington, D. C.

U.S. Department of Health, Education, and Welfare

1971 The Negotiated Contracting Process: A Guide for Project Officers. Office of the Secretary, Washington, D. C.

U.S. Department of Health, Education, and Welfare

1975 Contracting for Surveys in the Health Services Administration. Health Services Administration, Office of Planning, Evaluation and Legislation, Washington, D. C.

Joint Ad Hoc Committee on Government Statistics

1976 "Report of the Joint Ad Hoc Committee on Government Statistics", Statistical Reporter (September):301-310.

Federal Paperwork Commission

1977 Final Report of the Value/Burden Study Group (Circulating Draft). Part 4, C "An Assessment of Federal Survey Practices Done Under Contract and Grant". Value/Burden Study Group, Washington, D. C., IV-21 to IV-86.

Exhibit A - STANDARD CLAUSE ON USE OF PROBABILITY SAMPLING

Unless otherwise specified in the statement of work, any offeror's response to this Request for Proposal (RFP) shall contain probability sampling methods for the selection of respondents or subjects for any survey or other study in which sampling from a defined population is required. Sampling plans and cost proposals shall be based on such methods. If, however, an offeror feels that a nonprobability sampling approach would be more efficient, he may elect to submit an alternate proposal in addition to the required proposal. The alternate proposal must be fully justified and contain a separate cost proposal. Any offeror not submitting a basic proposal utilizing probability sampling methods shall be considered nonresponsive to the RFP.

Exhibit B - A CHECKLIST FOR RFP's FOR CONTRACT SURVEYS

Introduction - An RFP for a contract survey typically consists of two main parts:

A. Instructions to offerors on how to prepare a proposal and submit it to the agency. This section is subdivided into:

1. General instructions, including a brief description of the purpose of the proposed contract and a description of the evaluation factors which will be used to rate the technical proposals. Frequently, a statement is included giving the expected man-years or man-hours of professional effort considered necessary for the project. This information is intended to assist offerors in preparing their proposals.

2. Technical proposal requirements. This section lists the kinds of information which each offeror is expected to provide in his technical proposal. Normally, the technical proposal of the offeror to whom the contract is awarded (with any changes made in the process of negotiation) is made a part of the actual contract.

3. Business management proposal instructions.

B. Contract provisions, including "scope of work" statement. The scope of work statement sets out the background, objectives and specifications for the survey operations to be performed by the contractor. The amount of detail in the specifications may vary from one RFP to another, depending on the desires and technical expertise of the issuer.

This checklist is not intended to be a complete set of instructions for preparing an RFP. The final responsibility for preparation of RFP's rests with the procurement staff. The purpose of this checklist is to call attention to the principal elements of survey design and practice that determine the quality and utility of the outcome, and to suggest appropriate ways of treating these elements in the RFP. The goal, as in any survey, is to maximize the amount of information per dollar spent, keeping in mind that information is a function of the amount of error in the data.

A basic decision - At the outset, it is necessary to choose between the two basic methods of payment - fixed price and cost plus fixed fee. This choice is a subject of controversy, especially between issuers and offerors. Most, but not all, ORS contracts for surveys have used fixed price. Without trying to have the last word in this controversy, it is suggested that the fixed price approach is best if the issuer has a pretty good idea of what he wants and its cost is reasonably predictable.

Checklist

A. Instructions to offerors. The following should be included

1. Purpose of the survey

a. General statement of survey objectives

b. Are substantive results intended to be definitive, or is survey intended as a pilot test, or feasibility study?

c. Is survey descriptive or analytic?

d. How will results be used?

2. Information to assist bidders

a. Sampling frames, if any, available from agency.

b. Information on sampling and nonsampling errors obtained in similar surveys.

c. Agency policy on taping and other methods of monitoring interviews.

d. Whether use of government franked envelopes will be permitted for survey mailings.

e. Information on the contractor selection process, including a list of the selection factors to be used and their respective weights.

3. List of items that must be covered in offeror's technical proposal*

a. Detailed description of proposed sample design, including:

(1) Sampling frame

(2) Sample selection procedures

(3) Estimation procedure

(4) Procedure for estimation of variances

b. Data collection procedures

(1) Principal collection method(s)--face-to-face interviews, telephone interviews, mail questionnaires, other--with justification for method selected, especially in terms of expected quality of response.

(2) Procedures for training interviewers (if applicable).

(3) Methods to be used to achieve target response rate (see item B, 2, a, (5)).

(4) Methods and techniques to be used for minimizing response errors, especially for items known to be difficult or sensitive.

(5) Plans for supervision of interviewers and validation of their work.

(6) Plans for review and any necessary followup of questionnaires turned in by interviewers or returned by mail.

c. Processing procedures (if applicable)

(1) Clerical, coding and editing procedures--pretesting, personnel, training, verification.

(2) Keying procedures--verification

(3) Computer edits

(4) Procedures for tabulation and analyses

d. Procedures for protecting rights of data subjects and respondents, and for safeguarding confidential information.

e. Information on facilities and past experience.

(1) How will contractor arrange for necessary interviewing staff?

(2) Location, experience of interviewing staff to be used for survey.

*If the issuer wishes to pre-specify some of these elements, they should be omitted here and covered in the scope of work statement. See Part B, 2.

(3) Data processing facilities. Is any of data processing to be subcontracted?

(4) Brief summary of results and identification of agency references for last three completed surveys and for other surveys similar to this one. Indicate minimum set of items to be reported for each survey.

f. Name and experience of proposed project director and other key personnel who will work on this survey, with amount of time to be spent and principal functions for each person.

B. Contract provisions

1. "Deliverables". These are items which must be delivered to and accepted by the government at specified times.** Consider each of the following as a possible deliverable:

a. Periodic progress reports.

b. Draft questionnaire(s)

c. Proposed sample selection procedures.

d. Draft training materials and instructions for interviewers.

e. A report on pretest findings.

f. Draft specifications and instructions for data processing operations.

g. A specified number of copies of all final questionnaires, forms, instruction manuals, training materials, processing specifications, and other documents used in the survey operations.

h. A full accounting of the data collection results for the initial sample, with the following breakdown:

(1) Cases determined to be eligible

(a) Completed interviews
(b) Incomplete, by reason

(2) Cases determined to be ineligible, by reason

(3) Cases for which eligibility was not determined

i. Results of validation of interviews

j. Information on error rates found in verification of coding and keying operations.

k. Edited data tapes. If individual identifiers are needed (e.g., to merge survey and SSA program data) this should be specified.

l. Tabulations

m. Estimates of sampling error

n. Final report, including analysis of results and full report of survey operations, to the extent not covered by other items.

**Delivery dates should be specified in terms of time elapsed after award of contract. It may be desirable to have contingency provisions to allow for possible delays in agency or OMB clearances.

2. "Scope of work" provisions. These are specifications which the offeror must follow

a. Minimum set (should be included in all RFP's)

(1) A clear and complete definition of the survey population, including specifications of reporting units (e.g. individuals, households, beneficiary units) and of geographic coverage.

(2) Kinds of information to be collected, including specification of key variables.

(3) Required use of probability sampling at all stages of selection, and right of agency to review selection procedures.

(4) Level of reliability (sampling error) required for one or more key statistics. These requirements must be compatible with funds available for the survey.

(5) Target response rate. The term "response rate" should be clearly defined, including what is meant by a "completed questionnaire".

b. Optional items (may be included if considered appropriate)

(1) Requirements for pretesting.

(2) Acceptable data collection procedures. For example, for some purposes, mail questionnaires may not be considered acceptable. However, such restrictions should not be imposed unless there is good evidence to support them.

(3) Use of specific sampling frames and sampling selection procedures.

(4) Draft questionnaire(s). This will be helpful to offerors in estimating data collection and processing costs.

C. Some things to avoid in RFP's

1. Incomplete specifications

a. A sample of 1,000 persons. Does this mean 1,000 completed interviews or an initial sample of 1,000?

b. Estimates with a coefficient of variation of 5 percent. Which variables are subject to this requirement?

2. Over-specification. See item B, 2, a, (4). If the budget and the level of reliability are both specified, the budget should be large enough to achieve the desired level of reliability without cutting corners on other design features that affect the overall quality of the results.

3. Unnecessary constraints on survey design. Specific collection and processing procedures should neither be required nor ruled out unless there is objective evidence for doing so. Survey organizations should be allowed to demonstrate their expertise and ingenuity in developing the technical proposal.

SOME LESSONS LEARNED FROM CONDUCTING FEDERALLY SPONSORED SURVEYS

Eugene P. Ericksen, Institute for Survey Research, Temple University

1. The Situation

The job of responding to federal contracts for statistical surveys is fraught with ambiguity and frustration. This is because there is no clear standard for the quality of data and one has to play a guessing game about which standards will be used in judging a proposal or final report. Will they be standards of data quality, standards of policy relevance, or is the agency simply interested in getting a study done for the cheapest possible cost? Caught between the Scylla of poor quality research done for a small budget and the Charbydis of high quality research done at a price no one can afford, the result all too often turns out to be that the quality of the research is poor and the budget is exceeded. Given the importance of research, the large amount of money actually spent, and the large number of qualified statisticians, precisely how this occurs is a topic ripe for investigation by a student of organizational processes. It is also a topic of immediate concern for statisticians, since the quality of our collective product does little good for the legitimacy of our field.

I suspect that one basic cause has to do with the multiplicity of desirable surveys, which results in a budget for each that is insufficient for proper data collection. Why the number of surveys can't be reduced, with the additional money available from this reduction transferred to improve the quality of the remainder, probably has to do with the large number of agencies who need research done. Many of these agencies have insufficient budgets to commission quality surveys, and they seem to be reluctant to pool their resources. Nevertheless, there are many situations where budgets could be sufficient for quality research, but money is not spent wisely. As statisticians, we can have little impact on how decisions are made on which topics to carry out government research. However, there are aspects of the problem where I think we could fruitfully bring our influence to bear.

I would like to suggest that we should try to make progress toward solving two knotty problems. One is the general lack of agreement on standards and the other is the lack of objective criteria for making statistical choices. These problems were made particularly clear to me as a member of the Review Committee for the ASA Project on the Assessment of Survey Practices. Faced with the problem of how to decide when a survey could be judged as having met its objectives, we found it very difficult to write down a set of criteria. How does one compare a clustered sample for which a 65 percent completion rate was obtained and for which sampling errors were properly computed, with a clustered sample for which an 85 percent completion rate was obtained and sampling errors were not computed? This judgment becomes even more difficult when other issues are taken into account. For example, we had to make value judgments about the importance of validating interviews, the extensiveness of checking for data reduction errors, the quality of interviewer training, and the assessment of measurement error.

It is likely that most statisticians would agree that quality is paramount and therefore probability sampling should be used, sampling errors should be computed, interviews validated, data reduction checked, interviewers trained well, and that some check on the reliability or validity of data should be made. Unfortunately, the budgets of most government agencies writing survey specifications are not large enough that all these things can be done, and we lack a methodology of choice among criteria. Moreover, there are at least two issues which divide statisticians on defining proper practice. One is the proper method of computing a response rate and the other is the advisability of cluster sampling.

Most survey organizations report a response rate as the completion rate, the number of eligible respondents interviewed divided by the number of eligible respondents contacted. In spite of generally declining completion rates, this method of reporting a response rate can often produce a pleasant result, legitimately in the 85 to 90 percent range or higher. Unfortunately, nonresponse is often dominated by noncoverage, i.e., eligible respondents actually in the sample who are not contacted by interviewers. I would like to argue that the one proper way of computing a response rate is to obtain an independent estimate of the size of the universe and then compare this estimate to the weighted sum of eligible respondents, where the weights are equal to the inverses of the respective probabilities of selection. The ratio of the weighted sum to the independent estimate is the "true" completion rate which takes into account not only refusals but also households or telephone numbers where no one was contacted, incomplete enumeration of sample households, willful concealment of refusals on the part of interviewers, and sampling units not covered by the survey process. This includes housing units missed in the housing unit listing process in an area sample and housing units without telephones in a telephone survey.

The CPS appears to be one sample survey where this comparison is consistently done, and weights are computed to adjust for differential rates of nonresponse by various demographic subgroups. There appears to be no other survey organization which consistently makes this comparison and the typical method of reporting completion rates is to use the number of eligible respondents contacted as the denominator. Emphasis on this ratio encourages fudging, because an eligible respondent who is missed by an interviewer does not count the same as one who refuses to be interviewed. Emphasis on this ratio also

favors the use of quota sampling and random digit dialing telephone surveys because of the lack of concern for those who are missed by the survey process altogether. I suspect that one of the reasons the use of this procedure is continued is that it makes survey organizations look better and therefore increases their competitiveness. Estimates of total noncoverage are often embarrassingly high, and omitting such estimates significantly reduces the amount of explaining necessary to give to granting agencies. If granting organizations specified the size of the universe under study and insisted that this estimate be computed, the controversy over the proper computation of response rates could be ended.

In my opinion, there is a second area of more legitimate controversy. This concerns the ascendancy of cluster sampling and attempting to cover the entire population versus simple random sampling and not attempting to cover the entire population. On the one hand, it is typically impossible cost-wise to cover the entire household population of the United States without using some form of cluster sampling. Unfortunately, statisticians are increasingly using modern forms of multivariate analysis including log-linear modeling and logistic regression for which the error structure is not known when cluster sampling is used. Thus, some argue, it is impossible to make suitable inferences to the universe under study when we don't know how to compute sampling errors. Continuing their point, it is better to use a survey procedure such as random digit dialing or a mail survey where simple random sampling is possible, even though we know that part of the population is not being covered. Then proper statistical inferences can be made concerning the population that is covered and more speculative inferences can be made for the remainder. Given this hard choice, the added difficulty of choosing among features which all statisticians value makes the selection of a contractor from a set of competitive bids all the more difficult.

2. Organizational Factors Which Make the Problem Worse

These disagreements among statisticians weaken the basis on which rational decisions can be made by government agencies trying to decide on which survey organization to award a contract to. This decision-making process is weakened even further by two additional complications: (1) sampling theory is lacking which would aid in the choice among plans emphasizing different features of high quality research, and (2) choices about which features are most important to emphasize are not made by the government agency, either before or after the contract is awarded. Budget criteria make the final decision, and the result is that the completed research often has many unattractive features. Moreover, when the government organization isn't sure what it wants, prospective bidders are left to play a guessing game. I suspect that this ambivalence could be lessened by the more active participation of survey statisticians in the drawing up and writing of specifications for a proposed study.

Statistical procedures such as optimal allocation make it possible to balance a given reduction in variance against the corresponding increase in cost and to obtain a minimal variance sampling plan for a fixed cost or a minimal cost plan for a fixed variance. Unfortunately, sampling variation can be dominated by other sources of survey error due to unreliable or invalid measurement, noncoverage of important demographic subgroups, or sloppy data reduction procedures. We have no objective procedures for deciding on the optimal number of callbacks, or for estimating the number of questions needed to reduce measurement error for an important concept that is difficult to measure on a questionnaire. We cannot place dollar values on the personal training of interviewers relative to training by phone or through the mail. Similarly, we cannot place a dollar value on the validation of interviews. Given the disproportionate advances in sampling theory in the direction of estimating sampling errors, we lack objective criteria for assessing other trade-offs. For example, how does one compare a plan by which extra callbacks increase the completion rate by 5 percent, personal training reduces the unreliability of measurement by 10 percent, the validation of interviews weeds out the 3 percent of interviewers who cheat, and more careful editing procedures improve the reliability of measurement by 5 percent, against a plan which does none of these things but which uses optimal allocation to reduce variance by 10 percent for the same cost. These comparisons are not easy to make, even for an experienced, sophisticated statistician. Beyond measures taken to improve the bidding process, a priority area for statistical research would be to improve the methodology for assessing these tradeoffs.

In the meantime, hard choices must usually be made, and it appears that the choices are made all too often by administrators or financial officers who don't have the experience or know-how to properly confront these choices. Worse, the choices are usually not confronted until prospective contractors have submitted bids, which makes it extremely difficult for bidders to submit responsive proposals.

3. Suggestions for Improving the Bidding Process

I would like to suggest that three steps could be taken by government agencies to improve the process by which proposals are requested and selected for statistical surveys. These are (1) to make greater use of statisticians in drawing up and writing specifications, (2) to confront some of the difficult choices on survey specifications in advance and to indicate which choices have already been made and which choices they would still like to hear arguments on, and (3) to make greater use of statisticians to evaluate the collection and analysis of data after the project has been completed.

Most requests for proposals that we receive at the Institute for Survey Research give no indication about whether sampling errors should be computed, whether or not the granting agency is willing to pay for the validation of interviews and the personal training of interviewers, whether it is willing to pay for repeated measurements to evaluate the reliability of questionnaire items, whether or not substitutions should be permitted, or what kind of coverage rate is desired. A preference for probability sampling is usually assumed, and a specified response rate is sometimes given. Many of these choices could be made before the proposal specifications are written.

The present situation puts prospective contractors in a bind. Because of the standards we would like to set for ourselves, we prefer to compute sampling errors, to train interviewers in person, validate the majority of our interviews, use rigorous checking procedures in data reduction, and to collect repeated measurements to assess the reliability of our data. In fact, we insist on many of these features in our proposals, often with a religious fervor as "keepers of proper statistical practices." We have sadly lost many contracts to cheaper bidders because of this insistence on standards. The situation which often results is that the government agency is most willing to compromise on the computation of sampling errors or the assessment of measurement error. This is even more true when we subcontract for the collection of survey data to an organization which will take responsibility for analysis. Because it costs money to compute sampling errors, and because they, along with estimates of the extent of measurement error, make it more complicated to analyze data, we are often told not to compute sampling errors and assess measurement error. As a business in a highly competitive industry, we cannot afford to turn work away which fails to meet our "moral" standards, yet we are partially culpable for the poor statistical quality of some of the results. Because we find that we would confuse our interviewers and coders by relaxing our vigilance with respect to validation, training, editing interviews, and checking the accuracy of coding, the part of the survey process where we save money is in the assessment of sampling and measurement errors.

If statisticians were more intimately involved in the drawing up of survey specifications, it is likely that the hard choices would be faced in advance, and that the results of these choices could be included in the "Request for Proposals." If the specifications were rigorous, this would limit the set of competing organizations to those with the expertise to deliver the product. If the specifications were indicated to be less rigorous, organizations emphasizing high standards of research could choose not to bid. It would also be very helpful if a group of government statisticians, perhaps under the auspices of the Office of Management and Budget, got together to draw up a set of critical choices for survey specifications. Then, each RFP would have to state in advance its position on these choices, based on the amount of money available, the sample size necessary to provide useful information, and the minimum quality of information essential for intelligent decisions. The RFP would state whether sampling errors were desired, what the minimal coverage rate would be, whether or not interviews should be validated, and what type of interviewer training was necessary. It is likely that the forced confrontation of these choices would induce government agencies to opt for higher standards in order to justify the expenditure of money. This would strengthen the positions of contracting organizations and government researchers who emphasize high quality research and would likely improve the quality of research being done. If each RFP had to include a statement concerning whether or not sampling errors should be computed, it is likely that most proposals would include provisions for computing them.

Unfortunately, we know that survey statisticians and researchers in the government agencies do not have the final say concerning the choice of a survey organization. We at ISR have recently been in a situation where the research branch of an agency selected us to be the contractor, but the final decision was held in abeyance until the financial office had reviewed our budget and those of competing bidders to decide whether ours was truly cost efficient. How this was done in the absence of statisticians using statistical criteria is beyond me.

As a further check on practices, I suggest that funds should be put aside for the objective statistical evaluation of a study once it has been done. This evaluation would be public information, and would make it possible for the individuals and organizations doing the research to develop a "track record" which could be public information. For a survey with a total budget of several hundred thousand dollars, the cost of this evaluation would be a fraction of total costs. These reports would permit government organizations to check the past records of bidders.

It must be realized, however, that these procedures are likely to increase survey costs. As a result, if the standards of surveys are to be raised, a likely result is that fewer surveys would in fact be done. This could put some survey organizations out of business and result in a smaller volume of information available to government agencies. However, the quality of data would be higher and hopefully this would facilitate the decision-making process. I would argue that it is better to know you have a smaller amount of accurate information on an issue on which a decision is to be made, than to erroneously believe you have a large amount of accurate information.

RFP, GOVERNMENT, HOW TO LIVE WITH

Solomon Dutka, Audits & Surveys, Inc.

As with the other speakers on the program, my assignment is to discuss our experiences in coping with Federal Requests-for-Proposal and to suggest possible improvements which, I believe, would benefit the sponsoring agencies, the research contractors, and--ultimately--the citizen for whose benefit the research must be construed as having been undertaken and who must pay the bill.

The simplest approach to defining the roles of the Government agency and the research contractor is to try to compare them with the situation in private industry.

The research contractor, in general, attempts to play three roles: First, as an advisor on problem definition and methodological specification; second, as the executor of the research; and third, as the analyst who summarizes the survey findings.

In industry, in an increasing number of instances contracting companies will present the research agencies with carefully written specifications, reducing its first role. This is particularly true in the case of large corporations where internal research staffs may be large and available. The modal case, however, remains the one in which company research staffs are either small or too busy to try to do everything. In these cases, the representatives of the potential client and research company meet to discuss the problem and to try to reach a mutual understanding of the problem and the data needs. The research company then retires to prepare a detailed proposal which does the following:

1. Defines the problem;
2. Establishes that it has a competent grasp of the problem;
3. Outlines the methodological procedures for the study;
4. Establishes the technical, financial, logistic competence to get the job done.

The client company then finds itself with proposals from several bidding research companies, all of which may differ significantly in their design aspects and their estimated costs.

What is important and noteworthy in this situation is that it permits the research agency a great deal of flexibility in design and, in effect, encourages imaginative efforts in this area.

If the research company is assigned the study, it then has the additional responsibilities of executing the study, analyzing the data and writing a detailed summary of those findings. The research agency may even be invited into the corporate board rooms to discuss the findings and their implications with those who manage the company and must in some way implement the findings.

A final characteristic of survey research done for industry is its action orientation. In general, an existing problem motivates the research; the research, if successful, must provide guidance to the solution of the problem.

In dealings with the government, the situation is somewhat different. Taking the matter of project orientation first, we find that much of government sponsored research is policy-oriented rather than action-oriented. The deadlines for policy statements are often more slippery than those for actions and policy statements tend to be made on a more general level than action decisions. But that's only part of the difference between industry and government sponsored research.

In our experience, dealings with government agencies are initiated by an RFP. This is usually a very formal document in which the instructions on how to respond and the legal responsibilities of doing business with the government usually overwhelm the Statement of Work. The dealings with the government agency issuing the RFP are conducted at beyond arm's length. For example, all questions are to be directed to the Contract Officer who is usually not equipped to discuss any technical matters of the study design, analysis, etc. Ultimately, such questions are answered by a Technical Officer, via the Contract Officer, and sometimes even before the response to the RFP is due. When the contracting agency thinks there is a need for it, there may even be a briefing meeting to which bidders are invited. These have served some useful purposes. On one occasion, at least, the barrage of questions from the bidders was so devastating that the study had to be delayed over a year to permit the RFP-issuing agency to regroup and rewrite the RFP.

The goal of objective and fair evaluation of all proposals is absolutely vital. But the steps taken to achieve the goal sometimes work to limit the effectiveness of the research ultimately conducted. For example, the RFPs generally invite questions, but the lag between asking the questions and receiving the responses which must be broadcast to all bidders eats up valuable proposal-writing time. The RFPs, although often quite explicit in procedures to be employed, recognize that other alternatives are available and invite presentation of those alternatives as well. The burden, then, is on the bidder to prepare several proposals, all with equal enthusiasm. It's hard to write in detail and with promotional ardor on a plan which the research agency feels is either inadequate for the task, too expensive, or just plain poor.

In addition, the detailed specification of

research procedures often acts as a straitjacket and limits the contribution a research contractor can make to designing and executing the most effective study possible (either in terms of minimizing error for fixed expenditure or minimizing cost for required error). On the other hand, experience has taught us to be wary of invitations to be innovative. A recent RFP requested bidders to "stretch the limits of their imagination" in designing a study. One response, however, was turned down because it was "too new, it hadn't been tested."

In another instance, in an obvious attempt to give guidance to those responding to the RFP, it was specified that results be reported "...with an expected sampling error of ± 3% at the 95% confidence level". But, 3% of what? The same RFP did not even clearly designate the eligible respondent.

Writing a proposal is, for the research agency, a dance to entice the shy contracting agency. But writing a proposal for a government agency often makes the research agency feel it is dancing in galoshes. For example, the statement of work of an RFP often includes a good discussion of the background of the problem. The RFP then goes on to request a restatement of the problem in the bidder's own words to demonstrate his understanding of the problem. A simple reproduction of the RFP's description is nonresponsive. It the RFP says, "You will count apples", the response cannot say, "We will count apples". Instead, to be responsive, one might say, "The research contractor will determine the number of units in the class of fleshy and usually rounded and red or edible pome of fruit of a tree (genus Malus) of the rose family". Having carefully translated a simple declarative statement of four words and five syllables into something that most people can't understand, we have demonstrated an "understanding" of the problem. Why not a simple attestation that the bidder understands the problem and then let the study design itself testify to that understanding?

A final point -- because responding to a Government RFP is basically an expensive operation, tying up considerable man-hours, it is very troubling to discover after the RFPs have all been submitted that the selection of the successful bidder has been held up because the study hasn't yet been funded.

The relative importance of each of the three parts of the contractor's enterprise--advisor, executor, analyst--of course varies from study to study, but there appears to be a growing tendency to reduce the roles of advisor and analyst and to increase the role of doer. That, in itself, is a disappointing trend. The interesting parts of research projects are in the planning and analyzing. The room for innovation is essentially here.

The comparisons below are made to illustrate the differences between government and commercial surveys at the risk of overstating those differences.

CHARACTERISTICS OF GOVERNMENT AND COMMERCIAL SURVEYS

PURPOSES

GOVERNMENT

ENUMERATIVE

- TO ESTIMATE POPULATION PARAMETERS (E.G., POPULATION, UNEMPLOYMENT, PRICE LEVEL).
- PURPOSES ARE NOT USUALLY FRAMED IN TERMS OF IMMEDIATE ACTIONS TO BE TAKEN AS A CONSEQUENCE OF THE RESEARCH.
- STRATEGIC GUIDANCE
- FACTUAL DATA
- HOUSEHOLD DATA

COMMERCIAL

ANALYTIC

- TO TEST HYPOTHESES; TO SEEK BEST ALTERNATIVES.
- CONSEQUENCES OF DECISION ARE USUALLY SEEN MORE IMMEDIATELY AND DIRECTLY: RISK EVALUATION IS PART OF RESEARCH DESIGN.
- TACTICAL GUIDANCE
- ATTITUDINAL DATA
- INDIVIDUAL DATA

SAMPLING

GOVERNMENT

ACCESS TO GOVERNMENT RECORDS (E.G., SOCIAL SECURITY, TAX ROLLS) FOR SAMPLING PURPOSES.

COMMERCIAL

INGENUITY IS OFTEN THE ONLY WAY TO CONSTRUCT GOOD SAMPLING FRAMES (E.G., AREA SAMPLING). ACCESS TO CUSTOMER LISTS.

SCHEDULING

GOVERNMENT

LONGER PERIODS FOR STUDY EXECUTION. SCHEDULES GOVERNED BY LONG-RANGE PLANNING NEEDS FOR INFORMATION.

COMMERCIAL

INFORMATION NEEDS ARISE FROM IMMEDIATE PROBLEMS. THEREFORE, TIGHT SCHEDULES AND "YESTERDAY" DEADLINES.

RESPONSE PROBLEMS

GOVERNMENT

GOVERNMENT SPONSORSHIP OFTEN IMPLIES FORCE OF LAW AND ENHANCES RESPONSE RATES. THIS CAN, AT TIMES, ACT NEGATIVELY TO AROUSE SUSPICIONS OF RESPONDENT.

COMMERCIAL

RESPONSE DEPENDS ON RESPONDENT'S GOOD WILL.

BUDGET

GOVERNMENT

DATA ARE USUALLY PUBLISHED; GOVERNMENT HAS MANY CLIENTS.

DESIGN PRINCIPLE: MINIMIZE COST TO DELIVER FIXED VARIANCE.

COMMERCIAL

DATA ARE USED INTERNALLY; AGENCY HAS ONE CLIENT.

MINIMIZE VARIANCE FOR FIXED COST.

In order not to leave the impression that all is difficult in dealing with Government agencies, there are RFPs that are well-written, there are attempts on the part of the writers to 'ballpark' the study's budget, there are even attempts to establish lists of qualified bidders from which to select research agencies for given projects. But these instances tend to be the exceptions, making dealing with the government an extremely costly and time-consuming operation.

The following suggestions are made on the basis of our general experience with the bidding operation and with the feeling that improvements in these areas would, as we stated in our opening remarks, help all parties concerned, the Government agencies that require information, the research agencies, and the citizen:

1. Invite the participation of research agencies in the planning stages of a study.

2. On complex projects, select a small number of qualified agencies and, if necessary, give each a contract to develop a competitive design proposal.

3. Provide the responding research agency with greater design initiatives; don't specify all the details of the survey in the RFP.

4. The sampling specifications of an RFP should be written (or, at least, reviewed) by a sampling statistician; equally, the response to the RFP should be reviewed by a similar individual.

5. Make briefing sessions a matter of course for all projects; limit the size of each briefing session, scheduling more than one, if necessary.

6. The RFP should announce the budget level for each study.

7. Reduce the 'boiler plate' of the RFP.

DISCUSSION

Joseph Waksberg, Westat, Inc.

It is interesting that the speakers at this session who should represent opposing viewpoints, for example, the point of view of the Government vs. the contractor or a profit-making organization vs. one presumably mainly interested in research, arrive at essentially the same conclusions. They agree on the fact that the present system is not very good, on the problems that exist, and have approximately the same suggestions on how to improve current practices. I do not have any major disagreements with any of the speakers. However, I suspect that the speakers are underestimating the complexity of the situation and are too optimistic about the ability to make general improvements in a vast Federal system.

At least one reason for this is that the three speakers, although having diverse kinds of affiliations, have one thing in common. They represent organizations that have highly competent and sophisticated technical skills; they are concerned with quality; they understand the trade-offs between quality and cost; and they are aware of the many factors in statistical studies that affect quality and can assess the impact of trade-offs in expending resources on different aspects of quality. Unfortunately, these technical skills do not exist uniformly, either in the Government or in contracting organizations. This is what makes it difficult to conceive of a general and simplified procedure for preparing RFP's and choosing among bidders.

If Tom Jabine were the typical representative of a Government agency and Gene Erickson and Sol Dutka were typical representatives of contractors, the proposal in the Jabine-Pigman paper to have RFP's clearly state the objectives and funds available and leave the details to prospective bidders would undoubtedly produce the best results for the Government. I have much less confidence in the ability of many other Government agencies to choose the best offer when bidders are given such wide latitude. Erickson has pointed out the paucity of information that exists to help in choosing between high response rate or large sample size, on the increases in variances arising from clustering, etc. For many Government agencies, I suspect it is better for them to specify the major parameters of a survey design than to face a bewildering set of offers, some emphasizing sample size, others high response rates, still others more intensive training and supervision, with the agency staff not really knowing how to assess the relative merits of the different proposals. Also, there are probably contracting organizations with competent operational staff and who can produce work of reasonable quality if a Government agency described the required tasks in some detail, although they might not have the technical capacity to produce the basic plans. I am not sure they should be squeezed out of the possibility of doing some of the Federal statistical work.

Several of the speakers have commented on the desirability of the Federal agencies involving survey statisticians more directly in the preparation of the RFP's and in the choice of contracting organizations. I believe this is really the heart of the matter. Until more technically qualified personnel are involved in the contracting process, I doubt that changes in specifications or rules will have much effect. I am not implying that all Government agencies contracting statistical work are lacking such staff, but it is a fact that many do.

Although I agree with the basic content of the papers presented here, there are few specific issues I would like to comment on. First let me raise a few questions on several points in Erickson's paper.

(1) Nonresponse: I don't believe it is good practice to combine nonresponse and lack of coverage in a single measure. There are a number of reasons for keeping them separate: (a) For many surveys coverage is not under the control of the survey manager whereas response is. A combined measure does not provide information on whether the contractor is doing a satisfactory job. (b) Sometimes substitution is used for nonresponse adjustment. The proportion of substituted cases can be considered a measure of nonresponse. It is confusing to attempt to include undercoverage in the same measure. (c) Independent figures are not always available.

I agree with Erickson that coverage problems may be as important as nonresponse. However, I would suggest that agencies require computation of both nonresponse and coverage ratios (when methods exist for estimating coverage), but that these should be reported separately. This, incidentally, is the Census's practice.

(2) Cluster Sampling: I'm surprised to hear there is a controversy on its use. I have not come across it. What I have found, however, is the difficulty of deciding on a reasonable segment size for a particular study, and

the lack of information to help in such discussions. This situation will not be improved unless a body of information on intraclass correlations is built up. Both Erickson and Jabine have pointed out how rare it is to see an RFP which requires computations of standard errors. I have not seen a single RFP that asks for an analysis of between and within cluster variances, although with modern computational methods this would require little additional effort. Contractors are understandably reluctant to propose such efforts since the additional cost could put them at a competitive disadvantage. If the Government agencies do not specify that such analyses are required, statisticians in and out of the Government will never be able to choose intelligently among alternative sample designs.

(3) Research: Calculations of intraclass correlations are only part of a body of methodological research needed to improve data collection procedures. It is shortsighted of Government agencies not to include provision for some methodological research in large statistical projects. There are some exceptions. NCHS has funded research studies in advance of major studies, and this occasionally occurs in other agencies, but such research is quite rare and tends to be specialized.

Let me turn now to the Jabine-Pigman paper.

(1) Level of quality needed: The Jabine-Pigman paper starts off with the assumption that the major problem in Government-sponsored work is lack of quality. Although I have no quarrel with this emphasis, there is another side of the coin that needs attention. Not all surveys need high quality work as is implied here, and in some cases it is likely the Government is paying more for quality than is justified by the analytic needs of the data.

The main issue I found missing in the discussion today concerns the quality of data needed for a particular study. Possibly the title of the session resulted in a concentration on the lack of quality. However, there are situations when higher quality is built into a survey than needed. This occurs, for example, in decisions to use personal rather than telephone interviews (to avoid the bias of excluding non-telephone households) or decisions to include high-cost areas such as Hawaii and Alaska in sampling frames. It would be useful to give some consideration to assessment of the quality actually needed, in relation to the expected uses of data for a particular survey.

(2) Providing offerers with data on available budget and survey objectives: This is suggested as a way of improving the selection process. Mr. Dutka recommends a similar approach. Knowing the budget available is certainly essential for an intelligent response to an RFP. Keeping it hidden helps neither the Government nor the bidders. RFP's frequently refer to a "level of effort", but it has always seemed foolish to me to engage in such circumlocutions rather than clearly stating the maximum amount of money available for a study.

Asking offerers to develop survey proposals based only on a description of survey objectives is a sensible proposal for the larger agencies, with reasonably competent technical staffs. As I indicated earlier, I am not sure how this would work for smaller agencies. Possibly OMB should explore the feasibility of some kind of centralized system for smaller agencies.

(3) Probability sampling: Explicity stating that probability sampling is expected, and that use of nonprobability methods need special justification is obviously an important improvement. However, the agencies should accept the fact that under some circumstances nonprobability methods are appropriate. It should be noted that the pilot study on survey practices carried out by the Subsection on Survey Research Methods of the ASA used a purposive sample of projects. I assume there was a good reason not to use a probability sample.

If probability sampling is listed as a specific requirement in RFP's, then we may need to be more careful of our definition of probability samples. Will deliberate exclusions from the frame disqualify some sample designs if the words "probability sample" are taken literally? Some typical exclusions are: Alaska and Hawaii, group quarters, non-telephone households if random-digit dialing is used. How about if a Federal agency wants a study in a few locations - one county, four metropolitan areas, etc. Do the areas have to be selected on a probability basis as well as the units within them? There may be some legal ramifications if definitions are not carefully stated.

After hearing the three papers presented, I would like to summarize my own recommendations for improvements in the contracting process. In approximately

priority order, they are as follows:
(1) RFP's should indicate the maximum funds available.
(2) Biders should be provided with flexibility to trade-off different factors affecting quality, e.g., sample size vs. response rate.
(3) A method should be found for involving survey statisticians in the writing of RFP's and the choice of contractors. This is particularly critical for the smaller agencies. Perhaps some type of pool can be established for statistical assistance.
(4) For large projects, more use should be made of RFP's requesting preliminary proposals only, with the Government paying for more detailed designs for the two or three best initial proposals.
(5) Uniform and standard definitions need to be established for such concepts as response rate, probability sampling, what constitutes acceptable primary sampling units, etc.
(6) Some part of the funds for large projects should be set aside for methodological research.

PROFESSIONAL STANDARDS OF STATISTICIANS IN STATE AND LOCAL GOVERNMENTS

Harry M. Rosenberg
University of North Carolina

Anders S. Lunde
University of North Carolina

INTRODUCTION

The use and the importance of statistics in our society are growing. Its impact is increasingly apparent in all aspects of our lives, in the private sector, in our great institutions of learning, in our technologies, in political forums, and in the machinery of government. In government, its influence is felt at all levels--the federal, the state, and local sectors. As greater reliance is placed on quantitative evidence as a basis for both understanding and for decision-making in an increasingly complex society, the burden on and the responsibility of the statistician--as producer, as custodian, and as interpreter--of this important social tool will continue to grow.

Viewing the statistician and his products in the long historical context of mankind's development enables one to better appreciate the factors that have enhanced his role and today increasingly draw him into the public forum. As Jean Gibbons wrote so eloquently a few years ago, the enhanced role of statistics in our lives today is associated with civilization's long effort to cultivate increasing rationality in human decision-making (18).

Society's growing reliance on statistical information requires that we continuously strive to effect a better fit between public needs and the skills that we as statisticians possess. This calls for constant professional self-scrutiny, in terms of education and training, in terms of communication with the public, and in terms of generating high levels of expectation for ourselves. These aspects of self-scrutiny are all subsumed under the broad rubric of "professional standards," an area to which the American Statistical Association has directed its attention for over 25 years.

STANDARDS FOR STATISTICIANS

For over 25 years the American Statistical Association has addressed issues related to statistical standards through a variety of organized activities. As early as 1952, an Ad Hoc Committee on Statistical Standards recommended to ASA President William Cochran that the Association should work toward developing an agreed upon set of statistical standards, both technical and ethical, which could provide guidance to individual statisticians, in terms of standards to which published statistical results should conform, and procedures to assure valid statistical results (3).

However, interest in these "standards" questions has waxed and waned over time. Appraisals of the feasibility of establishing professional standards for statisticians have differed widely, depending upon the appraisers, their approach to the problems, and the historic context of their inquiry.

In the years since statisticians in the United States mobilized organized efforts to address "standards"questions, we have come to appreciate the wide range of issues involved, some of which come to the fore, then recede, then reemerge--all reflecting social and other forces impinging on the profession.

Standards for Practitioners or for Products?

Discussions about statistical standards often distinguish between standards applied to statistical products--such as timeliness, validity, reliability, accuracy--and those applied to statisticians, that is, to their competence levels and to their professional behavior. While this distinction is useful, particularly with respect to strategies for improving the quality of the statistical enterprise, these aspects of statistical standards are integrally related to one another. High competence standards for statisticians, and commensurate training levels, are likely to yield professionals who will bring to their work more sophisticated tools and higher performance expectations than those with less training. On the other hand, strategically speaking, the technical demands and performance standards associated with the statistical system itself--including the incentives and resources provided for realizing them--may be essential ingredients for stimulating high quality statistical work and for instilling a sense of professionalism among practicing statisticians. Indeed, essential demand may be a necessary condition for eliciting an appropriate supply response.

Albert Mindlin has stressed that one way the Federal government can help elevate local statistical standards is to insist on a certain level of sophistication in its work. Mindlin recently expressed particular concern when the Federal government asked local areas to assume less rather than more responsibility for producing local population estimates, suggesting that this approach was "deleterious to professionalism of state and local statisticians" (10).

Standards of Competence

A further distinction that bears on professional standards for statisticians is that between standards related to competence and standards related to professional behavior and practice. The competence question subsumes the many issues associated with statistical training and education, to which the American Statistical Association has devoted much attention. Competence standards are also central to consideration of individual certification and institutional certification and institutional accreditation--questions that come up from time to time in connection with broad inquiries into professional standards These questions arose, for example, in the

deliberations of the ASA Task Force on Professional Standards in 1970 and 1971. Standards for professional behavior, in comparison, are directly related to consideration of ethical conduct and performance.

With respect to certification, the Task Force on Professional Standards made some inquiries into this area, but took no definitive position on it (9a). Much earlier, in the 1950's the ASA Ad Hoc Committee on Statistical Standards under the Chairmanship of the psychologist and statistician Rensis Likert considered development of professional standards as essential and as a "necessary step before any certification procedure for statisticians can be established". The issue of certification for statisticians was raised again in 1973 by J. Boen and H. Smith who recommended that ASA give consideration to "imposing a structure on the statistics profession by certifying some statisticians as qualified to do applied work" (11).

When the question of certification was also raised among mathematicians in the early 1970's, the Board of Governors of the American Mathematical Association received a report which, in its general discussion of salient issues, seems relevant to certification for statisticians. J. G. Harvey and M. W. Pownall, authors of the AMA report, discussed both the accreditation of institutions and the related question of individual certification (19). They noted that among the traditional fields of liberal education, chemistry is one of the few fields with an accreditation system. Virtually all the others with special accreditation systems are professionally-oriented. According to Harvey and Pownall, chemists assess that minimum institutional standards have raised the quality of education in chemistry. But the authors caution that such standards, by being prescriptive, may threaten smaller institutions, discourage educational experimentation, and may rigidify curricula. They suggest that certification of mathematicians, might be accepted as evidence of qualification, but they question whether a system of certification by examination could really be designed to give a reliable evaluation of the qualities that it would purport to measure.

Those who considered these matters in Lester Frankel's ASA Task Force on Statistical Standards recognized some of these pros and cons as well. Herbert Alfasso, for example, spoke out in favor of certification for statisticians, but he recognized concerns that a program of testing for statisticians in connection with certification could be educationally stifling by restricting curricula, especially in a rapidly growing field like statistics.

In commenting on the implications of certification for state and local statisticians, Kenneth Rainey recently observed that much of the strength of statistics as a profession derives from its auxiliary role in support of other fields such as planning, public administration, engineering, and the regular professions. He sees a need for professional statisticians whose speciality is related to statistical analysis in support of government activities: but he is concerned that these professionals not be allowed to become a "priest-craft" (7b).

There seems little likelihood that pressures for the certification of statisticians will be great in the immediate future, since these pressures appear to most often arise from conditions of excess supply. Harvey and Pownall noted that certification and accreditation can be used to limit both the number of supplying institutions, as well as the number of professionals. The field of statistics does not appear in imminent danger of reaching such a condition in the near future.

Ethical Standards

When the Ad Hod Committee on Statistical Standards met in the early 1950's, many other professional associations were also addressing questions of ethical issues. For example, the American Psychological Association had formulated a code of ethical conduct for the profession. In her description of ASA activities in this area, Jean Gibbons notes that interest was high in the early 1950's under Rensis Likert's leadership, but after a survey assessment of membership interest, these issues were dropped 1956 by the Association as a formal matter (18).

In her description of statisticians' concern with this area, Gibbons calls attention to a number of related papers that have appeared in British and American journals. But her own work perhaps is one of the most cogent arguments for the importance of these issues, at a time when statistics and statisticians assume an increasingly important role in our society.

More recently in testifying before the Congressional Hearings on Statistical Coordination, James Knowles stressed the importance of ethical standards for statisticians in the organization and the operation of the Federal statistical system. He noted that foremost among the requirements for a quality statistical system is public confidence in its ethical integrity. "That confidence will not flourish unless the system enjoys the respect and confidence of professional workers activitely using the data coming out of the system..." (32).

State and Local Standards

Another organized effort of the American Statistical Association concerns itself with professional standards of statisticians in state and local governments. While many of the issues of statistical standards are basically the same as those discussed earlier without reference to the specific governmental context, there are two factors that make a focus on state and local governments particularly challenging.

The first is that the dramatic expansion of the state and local sectors during the past 20 years has given them a "frontier" character, in terms of opportunities for innovation and improvement, relative to the Federal sector. The second consideration is that a focus on state and local governments provides an opportunity to deal explicitly with an important set of factors that influence public statistical activities at all levels of government, namely, intergovernmental statistical issues. These issues speak of how the quality of our statisticians and the quality of our statistical products are influenced by the

relationships that exist among the Federal, state, and local levels of government.

A focus on state and local government by the American Statistical Association represents recognition of the importance of these governmental sectors in terms of their unique attributes and problems.

A FOCUS ON STATE AND LOCAL GOVERNMENT

In 1960, state and local governments employed about six million persons, or 2.5 times as many as the Federal government; by 1974, state and local government employed 12 million persons, or four times as many as the federal government. While the state and local sectors have continued to expand in terms of employment since 1970, the size of the Federally-employed labor force had not grown at all during 1970-74 (31). Rapid growth of the state and local sectors since the mid-1960's reflects a set of principles articulated by the Federal government in the late 1960's which stressed a greater role and responsibility for state and local government in the treatment of national problems (17,33,37). A reflection of this was the growth in Federal outlays to states, which Ullman showed expanded to $30 billion in 1971, four times the amount in 1960 (29).

Expansion of these governmental sectors has been accompanied by a certain amount of stress and strain. The accretion of new roles and the creation of new intergovernmental structures has required entirely different sets of relations both within and between governments. Strains have also arisen because the shift in responsibilities to states and local areas from the Federal government has been imposed on many areas which did not heretofore possess either the infra-structure or the personnel capable of discharging them effectively.

New responsibilities in many cases have been added to structures that were already rather complex, since the states and local areas had preexisting responsibilities to their constituents. Because states and local areas have sensed that the complexity of their governments has not been fully appreciated by the Federal government, there have been a number of efforts in recent years to elucidate and enunciate governmental processes, particularly those of states directed mainly at a Federal-level audience. Recent reports sponsored by the Council of State Governments (14, 15) describe the diversity and complexity of state governments, particularly with respect to their unique central "planning" functions and processes, which have no apparent structural or administrative counterparts at the federal level.

One theme that runs through these reports is an appeal to the Federal government to ease the burden imposed on the states by the "confusing, contradictory, duplicative, and overlapping mass of requirements and definitions in planning and program guidelines". The reports note, further, how Congress and the Executive Branch depend on state and local governments for program design and management in many areas; but that a major burden results from lack of coordination in program activities at the federal level: "Each federal program makes its unique and often conflicting demands on state government in its prescriptions for eligibility, planning, organization, fund matching, and procedures, imposing enormous burdens in terms of management functions and coordination at the state and local levels (15).

The rapid growth of the state and local sectors, the burgeoning programmatic responsibilities, and the absence of adequate program coordination at the Federal level have had consequences for management and administration at the other levels of government. These are reflected in intergovernmental statistical relations, and in the characteristics of statistical activities in states and local areas. They are reflected most insistently in the repeated plea, from municipalities and states, in 1967 and in 1977 for "better statistical coordination" (1, 25, 28).

Cooperative Statistical Programs

For statistical activities, the increased emphasis on state and local roles has built on preexisting structures and principles on intergovernmental cooperation know generically as the "Federal-State-local cooperative statistical programs". The first two programs of this type were initiated in 1917, and are now know respectively as the Cooperative Employment, Hours, and Earnings System of the U.S. Department of Agriculture. The Statistical Policy Division of the U.S. Office of Management and Budget describes these cooperative systems of data collection as "federally-initiated or sponsored statistical programs in which State agencies participate in the collection, processing, or utilization of nationally standardized statistics. The cooperative systems are undertaken for the mutual benefit of the participants, involve multiple states, and contain data of a recurrent nature which is intended to have broad applicability" (33).

The cooperative systems are built on an early federal recognition of an important and legitimate role for states in a national statistical system which was articulated as early as 1934 (17), and recently in the 1971 Report of the President's Commission on Federal Statistics. In the 1971 Report, Morris Ullman noted some of the advantages of these systems for reducing reporting burden, eliminating duplication, effecting economies through joint operations, and implementing principles of comparability (29).

In several respects, cooperative statistical programs have significance for state and local statistical activities. Just in terms of resources and manpower, some of these programs account for an important proportion of statistical support at the state and local levels. The two oldest programs--that of the Department of Labor and that of the Department of Agriculture currently fund, fully or in part, over 400 field positions in each state. The DOL budget for these field positions is about $3 million per year; the Agriculture budget for field staff is several times that. Another dozen or so programs in such areas as health, education, and law enforcement are neither as well-established nor as well-endowed in terms of resources as the DOL and

Agriculture programs (12, 17).

In addition to providing funds to states and local areas, the cooperative programs have been important means for improving the quality of statistical activities at these levels as government, as Morris Ullman noted (29). Katherine Wallman, in her discussion of these programs, indicates that statistical standards are integral to the cooperative statistical activities. "In each of the Federal-State Cooperative Systems of Data Collection, some attempt has been made to prescribe the definitional, quality, and timeliness standards which should be followed in the reproduction of the required data by the participating State" (33).

Training and Education. A significant contribution of the cooperative programs to enhancing statistical quality and professional standards of statisticians at the state and local has been through their related training and educational activities. Again, these are most developed in the older, better-established programs, where, for example, field staff are systematically exposed to training through seminars, meetings, and conferences, and in which staff are encouraged to take advantage of in-service training opportunities.

The potential for Federal leadership in promoting state and local statistical standards through education and training was recognized early in the evolution of the cooperative statistical programs. It was emphasized by both Herbert Alfasso and Morris Ullman in the Report of the President's Commission on Federal Statistics, where a particular training program of the Federal government was singled out as a model. This is the Applied Statistics Training Institute (ASTI) of the National Center for Health Statistics, established in the mid-1960's to provide training and educational opportunities for those working in the health area. Because of the high quality of ASTI's program, it has since become an educational resource serving many of the cooperative programs, as well as other statisticians at all levels of government. In the President's Report, Alfasso and Ullman, drawing on the example of ASTI, call upon the Federal government to take the lead in establishing a basic training program "for state and local statistical personnel covering both data gathering and data use". They recommend that costs be shared by the Federal government and the states (1,29).

In the area of training and education for statisticians, the Federal government has yet to develop a coherent and comprehensive model that could speak to the in-service and the career development needs of statisticians at all levels, from that of apprentice to that of high-level statistical administrator. Such a program could serve as a useful paradigm, if developed, for state and local governments. A recent study by the Statistical Policy Division of the U.S. Office of Management and Budget described various elements of such a program, which elements have been implemented by different agencies at different times, but never in a really coordinated manner (35). Such a program for career development, along with a comprehensive training institute oriented to the in-service training needs of all levels of government, could be useful paradigms and resources for improving professional standards of statisticians in state and local government.

Opportunities for Improvement. If the cooperative statistical programs have been successful in upgrading the quality of statistics and statisticians in state and local governments, through standard setting, resource transfer, training, and information exchange, they still present opportunities for improvement. Katherine Wallman has noted that across programs, there are still wide differences in the specification of statistical standards, in enforcement of adherence, and in resources provided to state and local areas to participate in these cooperative programs. Most troublesome, Wallman notes, is the lack of coordination of standards and guidelines among the statistical programs of the many sponsoring agencies, at the Federal level. In the absence of needed information exchange and coordination at the Federal level, the Federal statistical system, insofar as it affects states and local areas, falls far short of its potential (33).

Statistical Coordination

Among the factors frequently cited as having a bearing on quality of statistics at the state level is that of "coordination". In the National Conferences on Comparative Statistics sponsored by the National Governors' Conference in 1966 and 1967, the need for the improved statistical coordination at the state level was emphasized. Herbert Alfasso described these efforts in the Report of the President's Commission on Federal Statistics, where he identified as the most significant recommendation to come out of those conferences, that "each state develop an agency to coordinate statistical activities within the state and to serve as a channel to the federal government and to other states" (1). A similar theme was echoed by Jay Tepper in his presentation on "Intergovernmental Data Issues" at the 1977 meeting on data co-sponsored by the National Governors' Conference and the Council of State Planning Agencies (25,28). A related recommendation was made in a recent paper by Katherine Wallman which calls for establishing a "focal point" in each state to "coordinate State-level input to the Federal level on cooperative system" (37).

Despite repeated calls for improved statistical coordination at the state level, and certainly at the Federal level (32), there are some who have questioned whether the benefits of central coordinating units will meet expectations and who have asked if there might not be important costs in terms of effective communication between state and counterpart Federal statistical agencies. Rita Zemach sees the theoretical attractiveness of a central statistical coordinating agency at the state level, but does not feel that such units are practical in large states.

Progress toward establishing central coordinating units at the state level since the 1966 National Governors' Conferences recommended them has been limited. Herbert Alfasso reported that about 13 states had established such offices as of 1968, but by 1977 there was not much evidence that earlier momentum had been sustained;

indeed, some of these offices have since been disbanded. Alfasso stressed that statistical coordination at the state level requires Federal leadership through "providing recognition, technical guidance, and other assistance" (1).

Central Statistical Services

The concept of a focal point for statistical coordination at the state level is sometimes confused with that of central statistical services. While the two concepts are related to statistical standards, broadly defined, they are quite different from one another. Coordination need not imply central services, nor the reverse.

Albert Mindlin is a leading proponent of central statistical services, particularly at the municipal level (10). Mindlin emphasizes that the scale of governmental operations and the supporting resources in many states, and at the local level, are often insufficient to justify hiring highly trained statistical specialists in any one program, which simply could not "afford" them; but a central statistical office could hire professional statisticians who could, he asserts, design and carry out authoritative sample surveys; conduct skillful statistical analyses; apply specialized and highly efficient mathematical techniques such as statistical quality control to the improvement of government operations; and provide technical advice and consultation on the design, conduct, and evaluation of innumerable management improvements.

From the point of view of elevating professional standard and the quality of statistical work in state and local government, the concept of central statistical services is a plausible and an appealing one. However, given the imperatives of government organization, which is built around functional and line programs, it is often difficult to sustain interest in and support for central services, in the absence of strong outside incentives. As in the case of establishing focal points for the coordination of state statistical activities, it would seem that strong Federal incentives and leadership would be necessary to induce states and local areas to adopt a model of central statistical services for which Mindlin has made such a cogent case.

RECENT ASA ACTIVITIES RELATED TO STATE AND LOCAL PROFESSIONAL STANDARDS

ASA interest in statistical standards as they related to state and local government was initially stimulated by the work of the Social Science Research Council (SSRC) ASA Committee on Statistical Training about ten years ago. Recognizing the role of the state and local governments in an expanding range of program activities, and recognizing further that those assigned to statistical tasks at these levels of government often had little background in the field, the SSRC Committee, chaired by Conrad Taueber, suggested that concerted efforts be undertaken to "develop standards for statisticians in governmental service, with special reference to the needs of State and Municipal Services" (27).

ASA Committee

As a result of the ASA Board recommendation, the Ad Hoc Committee on Professional Standards of State and Local Government Statistics was organized on July 1, 1973, under the chairmanship of Anders S. Lunde. During the following year, the Ad Hoc Committee prepared a comprehensive set of recommendations for a long-range plan, as well as two reports.

Recommendations

In its recommendations to the ASA Board of Directors, the Committee of Professional Standards of Statisticians in State and Local Governments distinguished between those actions that would be focussed directly at state and local government statistical activities and those that could take advantage of Federal sponsorship of some of those programs.

ASA and the Federal Government. Recognizing the manifest accomplishments as well as the potential of Federal-state cooperative statistical programs for enhancing the quality of statistical work and for improving the professional stature of statisticians at all levels of government, the ASA would work with the Federal government, through the Statistical Policy Division and through the individual sponsoring agencies of major cooperative statistical programs, to:

1. Encourage development of uniform professional standards,
2. Review the structure and activites of the cooperative statistical programs, with a view to enhancing their statistical standards,
3. Encourage the development of training institutes for statisticians at all levels of government along the lines of Applied Statistics Training Institute, of the National Center for Health Statistics, and the Management Science Training Program of Training, of the U.S. Civil Service Commission.

ASA and State and Local Government. Working closely with representatives of state and local government, the ASA Committee would:

1. Explore state and local experience with Offices of Statistical Coordination and central Offices of Statistical Services,
2. Encourage and support the development of training programs for statisticians, including in-service training, on-the-job training, career and continuing education, and academic training, available to statisticians at all levels of government,
3. Encourage exchange programs for statisticians with universities and, through the Intergovernmental Personnel Act (IPA), among levels of government.

ASA and Members of the Profession. The ASA Committee on Professional Standards of Statisticians in State and Local Government recognizes that the extent to which statisticians are effectively used at all levels of government depends upon a clear understanding and appreciation of their

capabilities to inform the government process with their skills. This is very much a matter of education directed to those managers and administrators in state government with whom statisticians interact on the job. To facilitate the educational process, the ASA Committee would take responsibility for:

1. Developing general guidelines for job descriptions of statisticians, based on knowledge of existing position descriptions at the state and local levels, as well as on understanding the processes by which job descriptions are developed and modified in response to changing technological conditions and changing roles of statisticians in government,
2. Develop publications aimed at acquainting government program managers and administrators with the contributions that statistical reasoning and applications can make to government programs. Such brochures could be aimed at specific functional areas of state and local government responsibility,
3. Establishing panels of statisticians available to assist state and local areas in auditing the functions and jobs of statisticians, with a view to bringing these into closer alignment with the recommended general guidelines for these job descriptions. The panels could also be available to comment on other aspects of state and local statistical operations, including organization, administration, and implementation of statistical programs.
4. Organize a number of conferences and seminars under ASA auspices at the national, regional, and state levels that would serve as forums for promulgating and discussing general guidelines for statistical job descriptions; for discussing other issues of general concern to statisticians working in state and local government; for information exchange about intergovernmental and intragovernmental statistical issues; and for enhancing the understanding of statisticians' roles and capabilities by program managers and administrators.

CONCLUDING OBSERVATIONS

For over a quarter of a century, the American Statistical Association has actively addressed many of the issues associated with professional standards of statisticians. The changing foci of ASA activities are a response to shifting membership concerns which, in turn, are dictated by the social, economic, governmental, and technological context in which we live. During this period, we have all been witness to extraordinary changes that have affected the roles and responsibilities of statisticians.

Technological developments in data processing and computing capabilities have been truly revolutionary. They have facilitated date manipulation to an extend previously unimaginable. In addition, they have added to the cadre of persons working with quantitative data an entirely new group whose skills are closely aligned with the new technology.

Accompanying the technological revolution and amplifying it has been a dramatic increase in educational achievement throughout the population, resulting in a more informed public and one far more appreciate of the uses of statistical information. An emphasis on high level statistical skills now informs virtually every graduate program; and the emphasis is percolating down through the educational system, making important inroads today at the secondary school level.

In government, the use of statistical methods has proliferated, supporting such areas as planning management, budgeting, program evaluation, and many aspects of administration. As statisticians' skills are increasingly sought in the public and private sector, so too are statisticians drawn increasingly close to environments in which advocay dominates--in politics and in litigation--subjecting statisticians to public pressures as never before.

With new demands and pressures, with enhanced visibility and stature, statisticians are forced into exercises of self-scrutiny, in which they must ask themselves about their adaptation to constantly changing circumstances. We have attempted to review some of the issues that statisticians have addressed over the past 25 years in this continuing self-scrutiny, emphasizing certain issues associated with new intergovernmental circumstances.

REFERENCES

1. Alfasso, Herbert, "Report on Intrastate Statistical Coordination," Federal Statistics: Report on the President's Commission, II (1971), pp. 187-217.
2. American Public Health Association, Committee on the Statistics Section, "Report on Education and Training of Statisticians for Health Agencies," American Journal of Public Health, 60 (August 1970), pp. 1530-1545.
3. American Statistical Association. Ad Hoc Committee on Statistical Standards, "Report of the Ad Hoc Committee on Statistical Standards," American Statistician, (June-July 1954), pp. 19-23.
4. American Statistical Association, Ad Hoc Committee on Subnational Statistics, Progress Reports. Mimeographed (1973).
5. American Statistical Association, Ad Hoc Committee on Professional Standards of State and Local Government Statistics, The Role of Statisticians in State and Local Government. Mimeographed report available from the Executive Director, ASA (1974).
6. American Statistical Association, Ad Hoc Committee on Professional Standards of State and Local Government Statistics, Statistical Positions in State Government, with Index of Positions for Selected States, Mimeographed report available from the Executive Director, ASA, (1974).
7. American Statistical Association, Committee on Professional Standards of State and Local Government Statistics, (a) A Proposal for the Advancement of Professional Standards of Statisticians in State and Local Government

(January 1976). (b) Committee correspondence.
8. American Statistical Association, Federal Statistics Users' Conference, and Committee on the Integrity of Federal Statistics, "Maintaining the Professional Integrity of Federal Statistics: Report of the Committee", American Statistician 27 (April 1973), pp. 58-67.
9. American Statistical Association, (a) "Standards Task Force Report for a Study of Future Goals of the American Statistical Association," summarized in American Statistician (October 1971) (b) Task Force correspondence.
10. American Statistical Association, Subcommittee on Liaison with the Federal Government, Committee on Professional Standards of State and Local Government Statistics, "Progress Report by the Subcommittee Chairman Albert Mindlin," (April 28, 1975).
11. Boen, J. and H. Smith, "Should Statisticians Be Certified?" American Statistician 29 (August 1975), pp. 113-114.
12. Cavanaugh, Frederick J., "The Perspective of the Federal Government on the Role of State Government in Demographic Activities: A Joint Governmental Cooperative Effort," in Perspectives on State Demographic Activities, Harry M. Rosenberg (Ed.), (Oak Ridge, Tennessee: Oak Ridge Associated Universities, forthcoming).
13. Conant, James, Science and the Modern World, New York, 1952.
14. Council of State Governments, Intergovernmental Policy Coordination, (Lexington, Kentucky: Council of State Governments, September 1976).
15. Council of State Governments, State Planning: New Roles in Hard Times, (Lexington, Kentucky: Council of State Governments, September, 1976).
16. Deming, W. Edwards, "Principles of Professional Statistical Practice," Annals of Mathematical Statistics, 36 (December 1965) pp. 1883-1990.
17. Duncan, Joseph W. and Katherine K. Wallman, "Regional Statistics and Federal-State Cooperation,: Paper presented at the annual meeting of the Association of university Business and Economic Research, Williamsburg, Virginia (October 1975).
18. Gibbons, Jean D., "A Question of Ethics," American Statistician 27 (April 1973), pp. 72-76.
19. Harvey, J.G. and M.W. Pownall (Eds.), "Mathematical Education: The Question of Accreditation and Certification," American Mathematical Monthly, 77 (August-September 1970), pp. 746-751.
20. Hauser, Philip M., "Statistics and Politics," American Statistician 27 (April 1973), pp. 68-71.
21. Hogg, Robert V., "On Statistical Education," American Statistician (June 1972), pp. 8-11.
22. Joint Ad Hoc Committee on Government Statistics, "Report of the Joint Ad Hoc Committee on Government Statistics," Statistical Reporter 76 (September 1976), pp. 301-11.
23. Lunde, Anders S., "The Applied Statistics Training Institute (ASTI)," Statistical Reporter 73-3 (September 1972), pp. 37-40.
24. Mindlin, Albert, "A Professional Statistical Service for Local Government," paper presented to the Intergovernmental Seminar on Federal Statistics for State and Local Government Use, U.S. Bureau of the Census, Washington, D.C. (March 1974).
25. National Governors'Conference and Council of State Planning Agencies, "Numbers and Decisions: States' Use of Socioeconomic Data: Summaries of Issues Sessions," Mimeographed report available from the Conference Executive Director, National Governors' Conference, Hall of the States, 444 North Capitol Street, Washington, D.C. (June 1977).
26. Pieters, Richard S., "Statistics in the High School Curriculum," American Statistician 30, (August 1976), pp. 134-139.
27. Taeuber, Conrad, Frederick Mosteller, and Paul Webbink, "New Council Committee on Statistical Training, SSRC Items 21 (December 1967), pp. 49-51, and "Memorandum to the SSRC Committee on Statistical Training" (August 12, 1968).
28. Tepper, Jay, "Intergovernmental Data Issues," paper presented at the National Governors' Conference meeting on States' Use of Socioeconomic Date, Alexandria, Virginia, June 27, 1977.
29. Ullman, Morris B., "Federal Government Involvement in Data Collection for Sub-nation-l Areas," Federal Statistics: Report of the President's Commission, (1971), pp. 121-182.
30. U.S. Civil Service Commission, Position Descriptions, Statistician Series (GS-1530-0), February 1961.
31. U.S. Department of Commerce, Statistical Abstract of the United States, 1975, Washington, D.C.: U.S. Government Printing Office, 1975.
32. U.S. House of Representatives, 94th Congress, Subcommittee on Census and Population, Committee on Post Office and Civil Service, Hearing on "Coordination of Statistics," Serial No. 94-83 (February 25, and 26, 1976).
33. U.S. Office of Management and Budget, Statistical Policy Division, "The Federal-State Cooperative Systems of Data Collection," Statistical Reporter 77, (November 1976), pp. 37-48.
34. U.S. Office of Management and Budget, Statistical Policy Division, "OMB's Role in Planning and Coordination of Federal Statistics," Statistical Reporter 76-11 (May 1976), pp. 205-209.
35. U.S. Office of Management and Budget, Statistical Policy Division, "Professional Staffing and Professional Staff Training," Statistical Reporter 77, (April 1977), pp. 268-281.
36. Wallis, W. Allen, "Statistics in Nonstatis--tical Contexts," American Statistician 30, (November 1976), pp. 159-164.
37. Wallman, Katherine K., "Getting It All Together: The Development of Appropriate Relationships Between Federal and State Governments for Statistical Programs"

paper presented at the annual meetings of the American Statistical Association, Boston, Massachusetts (August 1976).
38. Zemach, R. and T.R. Ervin, "Records and Statistics," Health Services Reports 88 (May 1973), pp. 436-441.

DISCUSSION

Albert Mindlin, District of Columbia Government

My comments on the paper by Drs. Rosenberg and Lunde will be restricted to a few major points which I feel need elaboration.

1. The paper refers several times to existing Federal-State-local cooperative programs, and recommends their extension and elaboration. These programs have been very helpful. They have been a primary focus of good quality statical work at the State and local levels, and are the most immediately receptive points of Federal contact for improving State and local professionalism. There is, however, a unique reason for these features which should be clearly understood. Every one of these cooperative programs, except the one on population estimates operated by the Census Bureau, is primarily Federally-funded. For example, it is my present understanding that of the 52 State programs on labor statistics funded by US-DOL, 36 are 100 percent Federally-funded, and in the remaining States the funding is primarily Federal. Certainly there is far more and better statistical output because of these programs, and hopefully they are educating and addicting State and local governments to good quality statistics. But whether their presence has stimulated increased professionalism and professional statistical positions in other functional areas funded by State and local governments is not clear.

The Federally funded local professional statisticians in these programs not only produce statistics required by the Federal Government based on designs and specifications developed in the parent Federal agency, but also do professional statistical work of primary use by the State or local government, and designed by themselves fitted to State and local needs. This is in considerable contrast to the non-Federally funded cooperative program of population estimating. In this case, the professional work is done 100 percent by the parent agency, with some exceptions. The role of the State and local staff is primarily to provide input. It is not clear that this kind of program is stimulating state and local demographic professionalism, since it does not provide professional statisticians working on State and local demographic issues. (However, as an important aside, this program is an important protection to the objectivity of population estimates by making them relatively free from State or local political pressures. The local demographer is of course far more exposed to political pressure than the Federal Government. This is no mean advantage.)

2. The paper speaks of the difficulty of defining the term "statistician," and how it means different things to different people. It also makes various recommendations for improving the quality of State and local statisticians by training programs, expanding the cooperative programs, encouraging uniform professional standards, and other means. It makes only passing reference to the naiveté of administrators and the possibility of developing brochures aimed at administrators about what statistics can do.

It is this last point, the naiveté of administrators, which in my judgment is a major problem, perhaps the paramount problem in upgrading State and local statistical professionalism. It must be clearly understood that there are practically no autonomous professional statistical services in State and local government. There are no parallels to such Federal agencies as Census Bureau, BLS, BEA, NCHS, NCES. With some exceptions in health, to the best of my knowledge most statistical positions are either individual positions or small units embedded in operating agencies with layers of Divisions and and Bureaus over them. A few are in central planning agencies. Most administrators are abysmally ignorant of the planning, programming and managerial benefits of professional statistical operations. They tend to think of a statistician simply as a data collector. Sample survey design, statistical quality control, statistical modeling, and other technics of modern statistics are essentially unknown or dimly known to the typical highway, police, budget, revenue, fire, school administrator. In accord with this ignorance of modern statistical methods, the primary mission of whatever statistical work is done tends to be data gathering rather than professional statistical applications.

There are individuals sprinkled through State and local government who are doing professional level statistical work. But they are usually not called statisticians. Some administrators sort of intuitively know that certain things should be done that are statistical, and in filling a job they seek a subject-trained applicant who also has some statistical training. That is why such persons are sometimes found in subject matter positions. But the administrator can seldom articulate that his need is for professional statistical help or even recognize it in those terms very clearly. Thus having the job labeled something else ("educational analyst," etc.) is actually a protection because it permits a salary level that it could not attain if it were labeled "statistician."

Approaching the problem from this perspective, I suggest that the level of statistical work in State and local government is not likely to improve substantially until the subject matter administrators who make budget allocations are educated to the usefulness of modern statistical methods.

How can we raise the level of understanding of Departmental administrators? One way, as stated briefly in the Rosenberg-Lunde paper, would be to develop a series of educational brochures, each containing brief case studies and examples of how professional statistics can improve management--"How Statistics Can Help the Fire Department," "How Professional Statisticians Can Help the Board of Education." Another way is

to design a series of seminars for administrators, perhaps one-half day each, in addition to seminars for statisticians which are suggested in the paper. These perhaps could be prepared cooperatively with subject-matter professional associations, and given at professional meetings or by a statistician on the administrator's home ground.

In this connection the paper mentions "that about 13 States had established (central statistical coordinating units) as of 1968, but by 1977 there was not much evidence that earlier momentum had been sustained; indeed, some of these offices have since been disbanded." In my opinion we would gain considerable insight by pursuing this matter, such as investigating why momentum has not been sustained and some earlier efforts abandoned. I wish to propose some hypotheses:

(a) We are all aware of the severe budget contractions of State and local government in recent years. As a rough generalization with numerous exceptions, when budgets contract staff functions tend to be affected more severely than line functions. The garbage has to be collected, the potholes filled, the schools run. "Coordination," "planning," "research," and other staff functions being less directly visible to the electorate, are the easiest to cut.

(b) In accord with remarks made above, even to the administrator the benefits of "statistical coordination" are not clear enough to save the function when budgets contract. Indeed, this poor administrator understanding of what professional statistics can do to improve planning and management is a large factor in making these "central coordination units" rather powerless and ineffective offices when they are set up. Without authority to impose conformance to standards or to go into operating agencies with professional technics, and without professional staff capable of doing this, it is not surprising that "central coordination units" cannot accomplish much.

3. Most of the recommendations made in the paper are aimed at raising the professionalism of the State and local statistician. With respect to this target population the recommendations are good. But I have suggested above that in my opinion this is not where the principal problem lies. If, due to naivete of administrators there are few State and local professional statisticians and what ones there are have very little authority to act, then improving professional skills is likely to have minimal effect on improving State and local statistics. I wish now to suggest another major source of poor State and local statistics, far more serious than low statistical professionalism. Regardless of the quality of statistical analysis, such work must deal with existing data. If the data do not exist, good statistical professionalism can theoretically design survey or other procedures to generate original data. In fact this is seldom done at the State and local levels because the funds necessary to generate original data, even on a sample basis, very seldom exist. As stated earlier, the Federal statistical agencies that generate original data primarily for statistical purposes have almost no counterpart in State and local governments. The overwhelming source of statistical information other than that generated by the Federal Government is operating programs. In theory there is a cornucopia of information in operating files. But from the point of view of the practicing statistician, sophisticated statistical know-how pales into insignificance compared to the frustration of working with local operating files. An adequate discussion of this matter is beyond the scope of these comments. Suffice it to say here that, for the purpose of coherence, we may divide the problems and limitations of operating files into three general categories.

(a) Operating file organization and content seldom match the statistical need. An operating file is designed primarily to serve a daily operating mission--getting out water bills or tax bills or welfare checks or payroll, or assigning police or fire trucks fast, processing license renewals or violations of various codes, or making property assessments. File organination and content for statistical analysis for community planning, or management improvement, or to enable use of one agency's records to improve another agency's operations, or any purpose other than the immediate primary mission of the agency, is secondary and usually ignored, indeed isn't even perceived. Two examples:

(1) Every property assessment file has a land use code. This code is usually designed solely to distinguish properties necessary to make an assessment. Land use is the single most important datum for physical planning, but the land use code designed by the assessors' office is usually so abbreviated as to be of very limited use to the physical planner.

(2) Every housing code violation file has a violations coding scheme. Housing code violations can theoretically be a fertile source of information on housing condition. But a coding scheme that does not distinguish, say, between big cracks and little cracks, that counts 25 cracked windows as 25 violations, etc., while adequate to enforce the housing violations code, is virtually useless to evaluate the condition of a structure.

(b) Lack of automation. Numerous files of a local government, potentially of great statistical value, are effectively inaccessible because they are manual. One cannot stratify, sort, select, screen or do much of anything with them in any realistic time or cost frame.

(c) Poor quality. The quality of many operating files is atrocious. They are riddled with omissions, duplications, errors, anomalies, inconsistencies, undefined terms. The quality is often good for those few items of critical importance to the primary operating mission, but deteriorates rapidly for less critical items. Yet it is these less critical items that are often the most important to the statistician.

In the first instance, improving the level of statistical professionalism is not likely to

improve the quality and usability of operating files. But it can have a substantial effect if the professional statistician is utilized to address this issue. He can design statistical quality control procedures, computer edit checks, mediate between the data collection activity of the source agency and the needs of central and other agency planner, e.g., modifying coding schemes to make a file more useful to other agencies or for statistical purposes. The District of Columbia has a central professional statistical service. One of its major functions is to maintain an inter-agency computerized information system. This system obtains diverse operating files, integrates them so that unit records match, and then selects items from each file to permit statistical surveys and analyses far beyond anything possible from the separate files. In the course of matching files we impose both internal and cross-file computer edit checks. Each year we turn up thousands of errors this way, research and correct them, and feed them back to the source agencies, and thereby steadily improve their file quality.

If statistical staff is utilized this way, the upgrading of statistical professionalism can be of direct immediate benefit to operating agencies, whose needs almost always have budgetary priority over such luxuries as statistical analysis, management and planning. I submit that presenting the matter in this perspective may be one of the most fruitful ways to generate executive support for statistical professionalism in State and local government.

---oOo---

DISCUSSION

Rita Zemach, Michigan Department of Public Health

STANDARDS

ASA efforts should place emphasis on standards for statistics in government--for statistical products, processes, and functions--rather than on standards for statisticians. Statistical work in government is not necessarily done by persons classified as statisticians. The statistical work that has the biggest impact, and where there is the greatest concern for appropriate use of data, is in administrative activity, program measurement, program reporting. These activities use most of the statistical and data resource. They are often not recognized as being statistical activities, and the people involved in carrying them out may not be statisticians. By focusing on standards for the product, and pointing out the problems that arise when data are summarized for decision-making purposes, the ASA could generate the recognition that there is a technical area requiring special skills.

ASA ACTIVITES AND THE FEDERAL GOVERNMENT

I am somewhat concerned about the view that has been presented of statistics in the federal government. It's true that the major statistical agencies in the federal government have very high standards, have many outstanding skilled statisticians, and can provide guidance and leadership. State and local governments, however, are serivce and regulatory agencies. Their major relationship with the federal government is with their counterpart programs, and not with the statistical agencies. Thus, the major impact that the federal government has on state and local statistics is through program reporting requirements, and through the persons responsible for the program reporting systems. Judging by some of these systems, and some of the requirements, I am not optimistic that standards in state and local government can be improved through the federal government interaction.

The OMB Statistical Policy Division has been working hard to improve statistical processes and the reporting of statistics within the federal government. If this internal improvement of statistics could impact on the program reporting requirements of state and local governments; if there could be evidence of good statistical principles and practices in this ongoing activity, the federal government might thereby provide in-service training to all levels of government.

With regard to cooperative systems: cooperative activity is essential, but the different cooperative systems should be examined more carefully, before looking to them as a mechanism for improving state and local statistics.

Recent discussions have lumped the cooperative systems together. Actually, the earlier ones and the newer ones might be quite different. The early systems seem to have specific programmatic reporting objectives, and so were designed to collect and process data with some good sense of how the data were to be used. In at least one of the newer systems, the Cooperative Health Statistics System, we are trying to develop a "general purpose," multiple-user system. There are data collection prescriptions, and data processing standards, but no standards that relate to the statistical product itself--that is, to summarization, presentation, analysis. Thus, I don't see that program contributing to statistical quality and professional statistical standards, since there is nothing in the federal-state contractual relationship that even requires a statistician in the state cooperative project. There is, of course, an indirect spin-off, since in many states there is an effort to use the added resource to provide statistical services for state and local purposes.

To summarize: cooperative data systems can be a good mechanism for improving state and local statistics, but the statistical objectives have to be part of the system.

ASA ACTIVITES AND STATE AND LOCAL GOVERNMENT

I would urge ASA to refrain from recommending administrative arrangements for statistical work and stick to guidance by objectives. I have never heard the details of the states that are supposed to have central statistical coordination in state government. In fact, I believe there are relatively few distinct statistical programs or units in state and local government.

I am opposed to the idea of an agency in state government serving as a channel to the federal government for all cooperative efforts. This would seem to insert another layer of bureaucracy into the communication lines. Coordination within a state is desirable, but it is important to maintain direct federal-state technical communication along programmatic lines.

It also would not be appropriate for ASA to make a blanket recommendation for central statistical services It's true that there can be some benefit from aggregating statistical resources to serve a number of programs. The problem is that this may remove the statistician just far enough from the program so that he or she is not integrally involved in the program's day-to-day priority needs. Statistical work in state and local government involves such things as required program reporting and proposal documentation--state and local government agencies don't generally do research, and there are very few who do surveys. I am not for or against centralization, but feel

that it cannot be recommended as an administrative arrangement without knowing about the particular setting, the agency structure, the size of the establishment, and the people involved. Therefore, ASA should not endorse centralization, except as one possible alternative.

I would favor a program which encourages the ASA members who are government statisticians within a state to begin meeting, exchanging experiences and trying to work collectively to improve statistical standards.

JOB DEFINITION IN STATE AND LOCAL GOVERNMENT

Job definition is a problem, not only because of the nature of the job definitions for statisticians, but also because statistical work is carried out by so many persons who are not classified as statisticians--even the design of major statistical activities may be carried out by non-statisticians. In developing job description guidelines, ASA might look at other job classifications that should include requirements for statistical training.

A further complication is the fact that statisticians who are advancing in their careers are expected to move into administrative roles.

It might be helpful to have an exchange of ideas on the issue of appropriate job descriptions, with the discussion to include experienced statisticians from state and local government, as well as administrators with some understanding of statistical activities. It must be realized, for example, that decisions in government are not made on the basis of statistical analysis; statistical information is but one of many forms of information that contribute to decisions. The highest priprity in any agency is to meet program reporting requirements of one form or another, and description rather than inference is the predominant application of data.

FINAL COMMENTS

What is the greatest source of influence on statistics in state and local government? I mentioned before that the primary federal-state relationship is through reporting requirements. Going a step further, the source of most reporting requirements, at any level of government, is in legislation--either in the laws themselves, or in regulations. It is illuminating to examine the details of recent legislation, and note how many of them have quite detailed prescriptions for extensive statistical reporting. I can't imagine that anyone actually looks at all the material that pours in as a result.

It is also illuminating to note that some of the legislative requirements are things that experienced statisticians would know are impossible--or at least a methodology has yet to be developed. They require measurement of the unmeasurable; ask for demonstration of relationships that may not exist; and have resulted in a vast proliferation of number collection and manipulation, often undertaken by non-statisticians.

Perhaps ASA could examine this issue, and, as an independent professional organization, work with legislative staff towards more realistic expectations of what statisticians in state and local and federal government can produce.

THE INTERFACE BETWEEN STATISTICAL METHODOLOGY AND STATISTICAL PRACTICE

Gary G. Koch, University of North Carolina, Chapel Hill

1. Introduction

This paper is concerned with philosophical issues which require consideration whenever statistical methodology is applied to data. For this purpose, attention is focused on certain essential questions which statisticians must address for their efforts to be more meaningful than misleading. These include:

1. distinction between study population and target population,
2. distinction between variables under study and concepts which they are operationally assumed to represent,
3. role of technical assumptions pertaining to research design, existing state of knowledge, and statistical framework in which study objectives are formulated.

These and other aspects of statistical practice share "context" as a common theme. Here, "context" represents a perspective for evaluating the validity of the use of a particular statistical method through the relationship of the interpretation of its results to the specific nature of individual applications. Further clarification of this point of view is given for such topics as variable scaling, variable selection, and model building as applied to observational data, experimental data, and population sample survey data. For this purpose, an outline format discussion is given for two examples.

ACKNOWLEDGMENTS

This research was in part supported through a Joint Statistical Agreement with Burroughs Wellcome Company. The author would like to thank Jean Harrison, Jean McKinney and Pat Peek for their conscientious typing of this manuscript.

REFERENCES

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. Discrete Multivariate Analysis (M.I.T. Press, 1975).

Grizzle, J.E., Starmer, C.F., Koch, G.G. (1969). Analysis of categorical data by linear models. Biometrics 25, 489-504.

Higgins, J.E., Koch, G.G. (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. International Statistical Review 45, 51-62.

Koch, G.G., Freeman, D.H.,Jr., Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review 43, 59-78.

Koch, G.G., Freeman, J.L., Lehnen, R.G. (1976). A general methodology for the analysis of ranked policy preference data. International Statistical Review 44, 1-28.

Koch, G.G., et al. (1976). The asymptotic covariance structure of estimated parameters from contingency table log-linear models. Proceedings of the 9th International Biometric Conference, 317-336.

Koch, G.G., et al. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 33, 133-158.

Landis, J.R., et al. (1977) Parcat: a computer program for testing average partial association in three-way contingency tables. 1977 Proceedings of ASA Statistical Computing Section.

Example 1: Observational Data from a Case History Record System

a. Source: Clarke, S.H. and Koch, G.G. (1976). The influence of income and other factors on whether criminal defendants go to prison, Law and Society Review Volume 11, pp. 57-92.

b. Subject Matter and Objectives: To study historically a sample of persons arrested for certain types of burglary and larceny and to evaluate the extent to which an active prison sentence outcome was related to variables pertaining to the defendant's demographic status, specific type of offense, prior arrest record, etc.

c. Sample Design: All persons who were arrested for burglary, breaking and entering, and larceny (excluding automobile thefts and thefts involving less than $5.00) in Mecklenburg County, North Carolina with prosecutions begun during 1971. There were 798 such persons and all of them are included in the sample. Thus, the sample here corresponds to a total population.

d. Target Population:
   i. Local inferences: The population of interest is the actual sampled population which is restricted in time and place to 1971 and Mecklenburg County, N.C.
   ii. Extended inferences: The super-population of all persons who have been, are, or eventually will be arrested for burglary, breaking and entering, and larceny regardless of time and place from which the sampled population can be hypothetically regarded as a stratified simple random

sample with the strata being the cells of the multi-way cross-classification of those demographic, offense type, prior arrest record, etc. variables which have a statistically important relationship with whether a defendant receives an active prison sentence or not. In this regard, it should be noted that such a hypothetical super-population may not exist in which case any extended inferences are meaningless from a practical point of view. Nevertheless, an awareness of the existence of a context where they may be appropriate is still of interest.

e. Variables Under Study:
   i. Prison sentence status (Yes, No)
   ii. Type of offense charged (Non-residential burglary: NRB, Residential burglary: RB, Felonious and misdemeanor larceny: LARC)
   iii. Prior arrests (None: 0, One or more: 1+)
   iv. Arrest promptness (Same day: S, Later day: L). This variable is regarded as a measure of strength of evidence since it seems reasonable to assume that arrests which occurred very soon after the offense would tend to be based on more specific evidence (as opposed to circumstantial evidence) than those which occurred later.
   v. Median income of census tract of residence (Less than $7,000: L, At least $7,000 or suburban residents with unclassified income in terms of this definition: H). This variable is regarded as a general measure of socio-economic status as opposed to specific earnings.
   vi. Other variables which were considered included age, race, sex, and employment. However, after (ii)-(v) were taken into account, these other variables did not have a statistically important relationship with prison sentence status.

f. Data Display: The data corresponding to the multiway cross-classification of offense x prior arrests x arrest promptness x income x prison sentence status are summarized in contingency table format in Table 1.

TABLE 1

BURGLARY-LARCENY DATA: MULTI-WAY CROSS-CLASSIFICATION OF OFFENSE x PRIOR ARRESTS x ARREST PROMPTNESS x INCOME x DEFENDANT'S PRISON SENTENCE STATUS

Offense	Prior Arrests	Arrest Promptness	Income	Defendant's Prison Status Yes	Defendant's Prison Status No	Observed Prison Proportion	Est. s.e.	Statistical Model X	Model Predicted Prison Proportion	Est. s.e.
NRB	1+	S	L	15	14	0.517	0.093	1 2 0	0.537	0.050
NRB	1+	S	H	4	11	0.267	0.114	1 1 0	0.304	0.025
NRB	1+	L	L	12	22	0.353	0.082	1 1 0	0.304	0.025
NRB	1+	L	H	11	20	0.355	0.086	1 1 0	0.304	0.025
NRB	0	S	L	7	5	0.583	0.142	1 2 0	0.537	0.050
NRB	0	S	H	3	8	0.273	0.134	1 1 0	0.304	0.025
NRB	0	L	L	6	12	0.333	0.111	1 1 0	0.304	0.025
NRB	0	L	H	1	13	0.071	0.069	1 0 0	0.072	0.013
RB	1+	S	L	10	20	0.333	0.086	1 1 0	0.304	0.025
RB	1+	S	H	1	4	0.200	0.179	1 1 0	0.304	0.025
RB	1+	L	L	15	36	0.294	0.064	1 1 0	0.304	0.025
RB	1+	L	H	4	32	0.111	0.052	1 0 0	0.072	0.013
RB	0	S	L	2	8	0.200	0.126	1 1 0	0.304	0.025
RB	0	S	H	1	4	0.200	0.179	1 1 0	0.304	0.025
RB	0	L	L	1	17	0.055	0.054	1 0 0	0.072	0.013
RB	0	L	H	1	19	0.050	0.049	1 0 0	0.072	0.013
LARC	1+	S	L	15	51	0.227	0.052	1 0 1	0.193	0.032
LARC	1+	S	H	5	38	0.116	0.049	1 0 0	0.072	0.013
LARC	1+	L	L	14	68	0.171	0.042	1 0 1	0.193	0.032
LARC	1+	L	H	3	53	0.054	0.030	1 0 0	0.072	0.013
LARC	0	S	L	2	24	0.077	0.052	1 0 0	0.072	0.013
LARC	0	S	H	6	66	0.083	0.033	1 0 0	0.072	0.013
LARC	0	L	L	5	53	0.086	0.037	1 0 0	0.072	0.013
LARC	0	L	H	3	53	0.054	0.030	1 0 0	0.072	0.013

g. Data Analysis Strategies:

i. Local inferences. The basic framework is the multiple hypergeometric Model 0 in Appendix 1 with respect to which the hypothesis of randomness is being tested within two-way tables with fixed marginals and within sets of two-way tables with fixed margins. Of course, in a strict sense, all of the frequency counts are fixed constants (as opposed to random variables) because of the historical nature of the data. On the other hand, one can argue that there is still interest in the hypothetical question of whether or not the observed distribution of prison sentence status is at random with respect to each of the arrest description variables under study (for both the entire population as well as for sub-populations which are based on the other variables which are not being tested).

ii. Extended inferences. If the results of the local inference analysis seem plausible with respect to existing knowledge or theory for the substantive subject matter field to which the conclusions of the study are to be directed, then it may be realistic to assume the existence of a potential super-population to which such conclusions can be extended. In this case, the basic framework for analysis is the product multinomial Model 1 in Appendix 2. Thus, the respective proportions of defendants receiving prison sentences are random variables, and the principal objective of analysis is the characterization of the variation among them through the fitting of regression models and the testing of various hypotheses involving their parameters.

h. Results

i. Local inferences. Pearson chi-square statistics Q_P for testing the significance of the relationship between prison sentence status and the arrest descriptor variables are shown below.

Prison x Offense	Prison x Prior Arrests	Prison x Arrest Promptness	Prison x Income
Q_P(D.F.=2) = 48.35	Q_P(D.F.=1) = 15.23	Q_P(D.F.=1) = 4.43	Q_P(D.F.=1) = 19.45

Thus, offense, prior arrests, and income are significantly (α=0.01) related to prison sentence status in a strong sense (either with or without adjustment for multiple comparisons via Bonferroni inequality methods). However, the relationship between arrest promptness and prison sentence is only significant (α=0.05) in the weak sense where multiple comparison issues are ignored. Thus, caution should be exercised with respect to the nature of conclusions concerning this relationship (unless it was the one of primary interest in which case the other relationships would only be investigated from a descriptive as opposed to an inferential point of view).

Since the first order relationship of prison status to some of the arrest descriptor variables may be strongly influenced by the relationship of such variables to each other, partial association tests become of interest. For example, if the population is partitioned into three sets corresponding to offense type, to what extent is prison sentence status significantly related to income within these respective sets (taken together as a whole). A valid test statistic for this hypothesis (if the sample sizes within each set are sufficiently large) is the sum of the Pearson chi-square statistics for prison sentence status vs. income for the three offense types. Since Q_{TP}(D.F.=3) = 17.70, this partial association relationship is significant (α=0.01) with multiple comparisons issues being ignored since this test is typically in a philosophically different class than the ones described previously (i.e., either this type or the previous type or some third type may be regarded as the tests of primary interest from an inferential point of view but not all simultaneously since the spirit underlying the use of multiple tests here is the descriptive demonstration of support for a conclusion from several different points of view as opposed to the search for "significance" in the midst of randomness). If this type of analysis is continued further, the partial association between prior arrest history and prison sentence status after adjustment for (the joint partition of) offense type and income is considered. Here, however, the Cochran-Mantel-Haenszel statistic for which D.F.=1 is used in order to direct statistical power at average partial association alternatives (i.e., the extent to which the direction of the relationship between prior arrest history and prison sentence status is consistent across the six offense type x income sub-populations even though some of their respective magnitudes may be small). Since Q_{CMH}(D.F.=1) = 8.00, this partial association is significant (α=0.01). Finally, the partial association of arrest promptness with prison sentence status after adjustment for (the joint partition of) offense type, income, and prior arrest history is significant (α=0.05) with Q_{CMH} = 5.42.

In summary, several different types of hypotheses can be investigated for the purpose of local inferences about certain types of relationships among variables in populations of observational data. However, since the randomness in the data is only induced through the consideration of hypotheses, other types of statistical analysis like the estimation of standard errors for observed proportions, measures of association, etc. and the construction of confidence intervals cannot be undertaken in this framework because the data do indeed correspond to a population rather than to a sample from a population.

ii. Extended inferences. Here, the weighted least squares methods described in Grizzle, Starmer, and Koch [1969] are used to investigate the nature of the variation among the probabilities of defendants receiving prison sentences for the respective super-sub-populations corresponding to the (offense type x prior arrest history x arrest promptness x income) cross-classification. Since no prior information is available concerning the specific structure of a statistical

model for characterizing such variation, the complete contingency table is partitioned into six modules on the basis of offense type and prior arrest history, the two most important variables from a substantive point of view. Separate analyses are then undertaken within each of these modules in a manner which is primarily oriented toward their individual features but also attempts to reflect descriptively any consistency among them. As a result, the following models are found to be appropriate for the six (offense type x prior arrest history) modules

Non-residential Burglary
One or more prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.335 \\ 0.182 \end{bmatrix}$$

Model Q(D.F.=1) = 2.91
Residual Q(D.F.=2) = 0.46

Residential Burglary
One or more prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.111 \\ 0.189 \end{bmatrix}$$

Model Q(D.F.=1) = 6.89
Residual Q(D.F.=2) = 0.47

Larceny
One or more prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.071 \\ 0.122 \end{bmatrix}$$

Model Q(D.F.=1) = 8.76
Residual Q(D.F.=2) = 1.92

Non-residential Burglary
No prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.068 \\ 0.251 \end{bmatrix}$$

Model Q(D.F.=1) = 12.22
Residual Q(D.F.=2) = 0.15

Residential Burglary
No prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0.052 \\ 0.148 \end{bmatrix}$$

Model Q(D.F.=1) = 1.82
Residual Q(D.F.=2) = 0.01

Larceny
No prior arrests

$$\underset{\sim}{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.073 \end{bmatrix}$$

Residual Q(D.F.=3) = 0.65

After noting certain similarities among the predicted values across the six modules, the six distinct models are then synthesized together to form the overall model shown in Table 1 by methods analogous to those used in Koch, Freeman and Lehnen [1976] and Higgins and Koch [1977]. This model provides a relatively complete characterization of the variation among the proportions of defendants receiving prison sentences since

Model Q(D.F.=2) = 81.70 Residual Q(D.F.=21) = 6.01

Thus, the corresponding predicted proportions in Table 1 which are based on it represent a useful descriptive summary of the relationship between prison sentence status and offense type, prior arrest history, arrest promptness, and income in the hypothetical super-population from which the data are presumed to have arisen and to which any inferences are directed.

As a final comment, it should be noted that the structure of this model suggests the presence of substantial interaction among the respective arrest descriptor variables since the nature of the relationships within modules varies across modules. This aspect of the analysis may seem troublesome because the significance of such interaction has not been demonstrated. However, for this investigation tests for such interaction are not of direct interest because they pertain more directly to model reduction strategies than to substantively important hypotheses. For this reason, instead of reflecting a type of conclusion in the usual sense, interaction here corresponds to a concept in terms of which conclusions are qualified; i.e., the structure of the model for one module is not necessarily forced on the others unless the relationships within them are clearly compatible, when considered separately in their own right rather than with respect to the results of a statistical testing procedure, which sometimes are relatively weak (because they are directed at residuals of the models).

i. Conclusions

i. Local inferences. There do exist statistically significant relationships between prison sentence status and offense type, prior arrest history, arrest promptness, and income in the historical sample population of arrests for burglary, breaking and entering, and larceny in Mecklenburg County, North Carolina with prosecutions begun during 1971.

ii. Extended inferences. Since the sampled population in this application is sufficiently narrow to be of very limited interest in its own right, it is potentially desirable to argue that its local inferences can be extrapolated to some larger target super-population, even though it may not be possible to specify directly its location in time and place. This point of view is supported by the fact that the results of analysis are plausible with respect to existing knowledge in the criminal justice area (i.e., those relationships which are found to be statistically important are also, for the most part, substantively meaningful in both direction and magnitude). Thus, the structure of the fitted model $\underset{\sim}{X}$ in Table 1 and its corresponding predicted values for the data from this investigation are considered to be of general descriptive interest with respect to prison sentence outcome for burglary and larceny arrests, subject to the fundamental caution that any conclusions which are based on them should be regarded as

inherently tentative until it receives further support by similarly oriented studies for other locations and time periods.

j. Other sources of examples for observational data from case history records include:
   i. Analyses of highway safety injury data from motor vehicle accident populations which are defined in terms of record files for specific states and time periods;
   ii. Analyses of medical or dental outcome data for patient populations which are defined in terms of record files for specific clinics and time periods;
   iii. Analyses of product performance or safety history for consumer populations which are defined in terms of record files for specific distributors and time periods.

k. Summary statement concerning methodological issues: The most critical consideration underlying the interpretation of the analysis of this type of data is the relevance of the sampled population to the target population. This issue applies equally strongly for when the sample under study is a sample of a case history record system, as opposed to a total population like the one discussed here. Data quality and certain technical aspects of their statistical behavior are also important, but can often be assumed to satisfy the required conditions, since the scope of such analyses is either hypothetical or restricted to the descriptive summary of an isolated population of operationally prepared records (as opposed to measured phenomena). In other words, observational data from case history records basically stand on their own within whatever specific framework evolves. Thus, the principal question of interest is whether or not their analysis can be interpreted more broadly.

Example 2: Experimental Design Data

a. Source: Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969). Analysis of categorical data by linear models, Biometrics, Volume 25, pp. 489-504.

b. Subject Matter and Objectives: To investigate the relationship between the severity of the "dumping syndrome," an undesirable sequela of surgery for duodenal ulcer, and the nature and extent of surgery for four different types of operations which involve the removal of different amounts of the stomach.

c. Experimental Design: A multi-clinic randomized clinical trial involving suitably eligible patients who were treated in four participating hospitals during approximately 1966-1968.

d. Target Population:
   i. Local inferences: The population of interest is the actual study population as defined by the protocol inclusion criteria, the 1966-1968 time period, and the four hospitals.
   ii. Extended inferences: The super-population of all persons who have been, are, or eventually will be treated by one of the four operations regardless of time and place, from which the study population can be hypothetically regarded as a stratified simple random sample with the strata being the cells of the multi-way cross-classification of hospital, operation type, and any relevant demographic or patient diagnostic variables which have a statistically important relationship with the severity of the dumping syndrome which a patient experiences.

e. Variables Under Study:
   i. Dumping syndrome severity (None: N, Slight: S, Moderate: M)
   ii. Operation (Drainage and vagotomy: 0, Antrectomy (25% resection) and vagotomy: 1, Hemigastrectomy (50% resection) and vagotomy: 2, and 75% resection: 3)
   iii. Hospital (Hospital 1, Hospital 2, Hospital 3, Hospital 4)

f. Data Display: The data corresponding to the multiway cross-classification of hospital x operation x dumping syndrome severity are summarized in contingency table format in Table 2.

g. Data Analysis Strategies:
   i. Local inferences. The basic framework is the multiple hypergeometric Model 0 in Appendix 1 with respect to which the hypothesis of randomness is being tested for the relationship between operation and dumping syndrome severity in the set of four two-way tables corresponding to the respective hospitals. Here, the marginal distributions of operation type are regarded as fixed in principle by the nature of the experimental design (but may actually be subject to some inherent random variability because of possible protocol violations, missing data, etc.). In addition, the marginal distribution of the dumping syndrome severity is regarded as fixed under the null hypothesis (because it implies that the dumping syndrome severity for each separate patient is not affected by the operation that is experienced and thus, its distribution remains the same for all realizations of the treatment randomization process). Otherwise, since the dumping syndrome severity data are ordinally scaled, the types of alternatives which are of primary interest are location shifts which are indicative of the extent to which the dumping syndrome tends to be more severe for certain operations than others. It is also of interest to investigate the extent to which these location shifts are related to the ordinal scaling of the operations with respect to the amount of stomach removed.
   ii. Extended inferences. If the local inference results indicate a significant difference which is considered to be generalizable to some larger population, then it becomes realistic to analyze the data in terms of the product multinomial Model 1 in Appendix 2. In this regard, the

TABLE 2

DUMPING SYNDROME DATA: MULTI-WAY CROSS-CLASSIFICATION OF HOSPITAL × OPERATION × DUMPING SYNDROME SEVERITY

Hospital	Operation	Dumping Syndrome Severity N	S	M	Normalized Uniform Average Score	Est. s.e.	Statistical Model $\underset{\sim}{X}$	Model Predicted Average Score	Est. s.e.
1	0	23	7	2	0.17	0.05	1 0	0.20	0.03
1	1	23	10	5	0.26	0.06	1 1	0.24	0.02
1	2	20	13	5	0.30	0.06	1 2	0.28	0.02
1	3	24	10	6	0.28	0.06	1 3	0.33	0.03
2	0	18	6	1	0.16	0.05	1 0	0.20	0.03
2	1	18	6	2	0.19	0.06	1 1	0.24	0.02
2	2	13	13	2	0.30	0.06	1 2	0.28	0.02
2	3	9	15	2	0.36	0.06	1 3	0.33	0.03
3	0	8	6	3	0.36	0.09	1 0	0.20	0.03
3	1	12	4	4	0.30	0.09	1 1	0.24	0.02
3	2	11	6	2	0.26	0.08	1 2	0.28	0.02
3	3	7	7	4	0.42	0.09	1 3	0.33	0.03
4	0	12	9	1	0.25	0.06	1 0	0.20	0.03
4	1	15	3	2	0.18	0.07	1 1	0.24	0.02
4	2	14	8	3	0.28	0.07	1 2	0.28	0.02
4	3	13	6	4	0.30	0.08	1 3	0.33	0.03

variation among certain mean scores is investigated through the fitting of regression models and the testing of various hypotheses involving their parameters.

h. Results

i. Local inferences. The Cochran-Mantel-Haenszel statistic is used to test the significance of the partial association between the operation type and the dumping syndrome severity after adjustment for (the partition of) hospital. Moreover, to target statistical power at order partial association alternatives (i.e., the extent to which the probability of more severe dumping syndrome outcomes tends to increase (or decrease) with the extent of the operation in terms of larger amounts of stomach removed), the correlation mode for this statistic with D.F.=1 is used. In this regard, two types of scores are potentially appropriate. The first is ridits (or equivalently rank scores) which provides a partial Spearman rank correlation analysis of the data. The principal advantage of this approach is that it provides a framework in which the potentially difficult question of variable scaling can be avoided. Its disadvantage is that its results do not necessarily have a straightforward interpretation with respect to such scales since the analysis proceeds in terms of an index. Alternatively, uniform (or normalized uniform) mean scores can be used. Here, the functions $(p_S + p_M)$ and p_M are regarded from a substantive point of view as equally important measures of dumping syndrome severity for the purpose of assessing variation among the operations (on a within hospital basis). Thus, if there is no variation among the operations, there is no variation with respect to each of these measures and hence no variation with respect to their sum $(0p_N + p_S + 2p_M)$. Alternatively, if there is variation with respect to their sum, there must be some variation with respect to either $p_S + p_M$ and/or p_M. Similarly, if F_0, F_1, F_2, F_3 are measures of dumping syndrome severity for operations 0, 1, 2, 3 respectively, then the ordered pairwise differences (F_1-F_0), (F_2-F_0), (F_3-F_0), (F_2-F_1), (F_3-F_1), (F_3-F_2) will all be expected to be null if there is no variation among the four operations. Thus, their sum $G = (-3F_0 - F_1 + F_2 + 3F_3)$ is also expected to be null. However, if this sum is concluded to be non-null, then there must be some variation among the four operations. Otherwise, it should be noted that the function G is constructed to compound pairwise differences which are arranged to reinforce one another if indeed the probability of more severe dumping syndrome outcomes does tend to increase (or decrease) with the extent of the operation. Thus, uniform scores can be used for both dumping syndrome severity and operation on the basis of statistical power arguments with respect to order association alternatives. In addition, although the scaling which they induce on the categories for these variables may not necessarily have a meaningful substantive interpretation, they do nevertheless provide a quantitative framework which can be used for descriptive statistical purposes.

For this example, the normalized uniform scores (0, 0.5, 1) will be used for the dumping syndrome so that "moderate" is regarded as the principal response level of interest and "slight" is interpreted as half-way between in the sense that two people with "slight" are considered to be equivalent for comparison purposes to one person with "none" and one person with "moderate." Since the operations are naturally scaled with respect to the amount of stomach removed, the scaling 0, 1, 2, 3 does not require any further explanation.

Finally $Q_{CMH}(D.F.=1) = 6.34$ for uniform scores and $Q_{CMH}(D.F.=1) = 6.92$ for ridit scores. The former is significant at ($\alpha=0.05$) and the latter is significant at ($\alpha=0.01$). Thus, both indicate a significant relationship between dumping syndrome severity and the extent of the operation. Moreover, it should be noted that the focus of these statistics on order association alternatives at the beginning is critical because the overall Cochran-Mantel-Haenszel statistic $Q_{CMH}(D.F.=6) = 10.60$ is not significant ($\alpha=0.10$).

In summary, the partial association between dumping syndrome severity and extent of operation can be investigated for the purpose of local inferences with respect to the set of patients defined by the protocol inclusion criteria, the 1966-1968 time period, and the four hospitals. For this purpose, the only assumption required is the validity of the randomization process by which patients were assigned to operations. However, once the hypothesis of randomness is rejected, the hypergeometric Model 0 in Appendix 1 is no longer applicable. In addition, since such rejection implies the existence of a significant relationship in a local inference sense, it then becomes of interest to characterize descriptively its nature in terms of fitted regression models for extended inference purposes with respect to some larger super-population.

ii. Extended inferences. As with Example 1, the weighted least squares methods in Grizzle, Starmer, and Koch are used to investigate the nature of the variation of the distribution of dumping syndrome severity for the respective super-sub-populations corresponding to the operation x hospital cross-classification. More specifically, attention will be focused on the mean score function $F = (0.5p_S + p_M)$ because of its sensitivity to location shifts and its compatibility with certain asymptotic (central limit theory) assumptions as discussed in Koch et al. [1977]. Secondly, since the participating hospitals all followed the same basic protocol with respect to the inclusion of eligible patients in the study, the conduct of the four operations, and the evaluation of patient response, it is reasonable to assume a priori that the variation among the mean score functions F_{hi} (where h = 1, 2, 3, 4 indexes hospitals and i = 0, 1, 2, 3 indexes operations) can be characterized in terms of an additive model with respect to hospital and operation effects. One formulation for such a model is

$$E\{F_{hi}\} = \sum_{k=1}^{t} \beta_k x_{hik} \quad \text{where} \quad \{x_{hi} = 1 \text{ for all } h, i\} ; \quad \left\{ \begin{array}{ll} x_{hi2} = 1 & \text{if } h = 2 \\ \quad\;\; 0 & \text{if } h = 1, 3, 4 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi3} = 1 & \text{if } h = 3 \\ \quad\;\; 0 & \text{if } h = 1, 2, 4 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi4} = 1 & \text{if } h = 4 \\ \quad\;\; 0 & \text{if } h = 1, 2, 3 \end{array} \right. ; \quad \left\{ \begin{array}{ll} x_{hi5} = 1 & \text{if } i = 1 \\ \quad\;\; 0 & \text{if } i = 0, 2, 3 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi6} = 1 & \text{if } i = 2 \\ \quad\;\; 0 & \text{if } i = 0, 1, 3 \end{array} \right. \quad \left\{ \begin{array}{ll} x_{hi7} = 1 & \text{if } i = 3 \\ \quad\;\; 0 & \text{if } i = 0, 1, 2 \end{array} \right.$$

in which case β_1 represents a predicted value for operation 0 in hospital 1; β_2, β_3, and β_4 represent incremental effects for hospitals 2, 3, 4 respectively; and β_5, β_6, and β_7 represent incremental effects for operations 1, 2, and 3 respectively. The appropriateness of this model is confirmed by the non-significance of its goodness of fit statistic $Q(D.F.=9) = 6.33$ (which here corresponds to the hospital x operation interaction). Thus, certain hypotheses with respect to the parameters of this model can be tested in order to identify whether or not further model simplication can be undertaken. In this regard, the following hypotheses are of interest:

Source of Variation	Hypothesis Formulation	D.F.	Q_C
Hospitals	$\beta_2 = \beta_3 = \beta_4 = 0$	3	2.33
Treatments	$\beta_5 = \beta_6 = \beta_7 = 0$	3	8.90
Equality of Treatment Increments	$(\beta_6 - 2\beta_5) = (\beta_7 - 3\beta_5) = 0$	2	0.30
Hospitals and Equality of Treatment Increments	$\beta_2 = \beta_3 = \beta_4 = 0$ $(\beta_6 - 2\beta_5) = (\beta_7 - 3\beta_5) = 0$	5	2.61

On the basis of these results, hospital effects can be removed from the model and treatment effects can be simplified to a single equal increment (linear) parameter. The specific structure of this model is shown in Table 2 together with corresponding predicted values and their standard errors. These predicted values indicate that the dumping syndrome severity functions $\{F_{hi}\}$ increase from the value of 0.20 for operation 0 to the value of 0.33 for operation 3 in increments of 0.04 per quarter of stomach removed (for each of the hospitals). Otherwise, the goodness of fit statistic for this model Q(D.F.=14) = 8.94 is non-significant (α=0.25) and the test statistic for the equal increment parameter Q(D.F.=1) = 8.98 is significant (α=0.01).

In summary, the structure of the model in Table 2 indicates that hospital effects can be ignored for the set of four hospitals which participated in this investigation during 1966-1968. Thus, it is plausible to extend this conclusion to all hospitals and all years and thereby argue that the equal increment relationship between the dumping syndrome severity functions $\{F_{hi}\}$ and extent of operation which was found to exist for these data could be generalized to this super-population.

i. Conclusions

i. Local inferences. There does exist a significant relationship between dumping syndrome severity and extent of operation after adjustment for (the partition of) hospital in the actual study population as defined in terms of the research design protocol, the four participating hospitals, and the 1966-1968 time period.

ii. Extended inferences. Since the study population in this application is relatively narrow in its definition, it is of interest to argue that its local inferences can be extrapolated to some larger target super-population of hospitals for which the same conclusions would be anticipated in future (or other) time periods. This point of view is supported by the non-significance of hospital effects and hospital x operation interaction. Otherwise, the extended inference analysis would have needed to take into account certain patient demographic and diagnostic covariables in order to produce a more complex super-population framework with respect to which hospital effects could be potentially ignored. In other words, if there is variation among hospitals, then it is not realistic to generalize local inferences for the self-selected (by their willingness and/or ability to participate) hospitals in this study to some larger population. However, if such variation can be statistically explained in terms of variation in patient populations with respect to certain covariables, then extended inferences are plausible for the stratified super-population corresponding to the multi-way cross-classification of these covariables and operation. Thus, from this type of point of view, the structure of the fitted model $\underset{\sim}{X}$ in Table 2 and its corresponding predicted values for the data from this investigation are considered to be of general descriptive interest with respect to the relationship between dumping syndrome severity and extent of operation. Otherwise, conclusions which are based on such results are subject to the same type of caution indicated for Example 1 in the sense of being inherently tentative until they receive further support by similarly designed studies or observed experience at other hospitals during future time periods. On the other hand, the practical importance of such qualifying statements is potentially reduced considerably for such experimental situations when they are conducted in a carefully controlled manner with strict adherence to the design protocol and strict maintenance of data quality control and when they include participating hospitals (or clinics) which reflect coverage of a broad range of patients in terms of geographic area, demographic characteristics, and diagnostic characteristics.

j. Other sources of examples for experimental data include

i. Experiments involving animals from certain types of breeding colonies;

ii. Experiments involving agricultural plots in certain judgmentally (as opposed to randomly) selected geographic areas;

iii. Experiments involving persons who are linked to certain institutions (schools, hospitals, criminal justice system) at certain judgmentally (as opposed to randomly) selected locations.

k. Summary statement concerning methodological issues. The most critical consideration underlying the interpretation of the analysis of experimental data is data quality as reflected by the extent to which there was strict adherence to the research design protocol. In this regard, potential sources of difficulty include protocol violations, missing data, and certain sources of measurement error (and/or bias). More specifically, if these types of data quality problems can be avoided (or managed in a substantively acceptable manner), then local inferences (concerning the randomized variable; e.g., operation type) can be undertaken in an assumption-free manner via the probabilistic structure (e.g., Model 0) induced on the data by the randomization component of the basic experimental design; and those concerning other variables can be undertaken in the hypothetical sense which was described with respect to Example 1. If there is interest in extending the scope of the conclusions of the local inference results, technical aspects of data analysis like variable scaling, variable selection, and model formulation (as discussed previously for this example and also in Koch, Freeman, and Lehnen [1976] and Higgins and Koch [1977]) become important so that the linkage between the sampled population and the target super-population is operationally defined in a sufficiently relevant manner.

Appendix 1: Model 0

Let h = 1, 2, ..., q index a set of (s×r) contingency tables. Let i = 1, 2, ..., s index a set of sub-populations which are to be compared with respect to a particular response variable for which the outcome categories are indexed by j = 1, 2, ..., r. Let n_{hij} denote the number of subjects (or study units) in the sample corresponding to the h-th table who are jointly classified as belonging to the i-th sub-population and the j-th response category. These frequency data can be summarized as shown in Table A1.

Table A1

Sub-population	Response Variable Categories 1	2	...	r	Total
1	n_{h11}	n_{h12}	...	n_{h1r}	$N_{h1.}$
2	n_{h21}	n_{h22}	...	n_{h2r}	$N_{h2.}$
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
s	n_{hs1}	n_{hs2}	...	n_{hsr}	$N_{hs.}$
Total	$N_{h.1}$	$N_{h.2}$	...	$N_{h.r}$	$N_{h..}$

In this framework, $N_{hi.} = \sum_{j=1}^{r} n_{hij}$ denotes the marginal total number of subjects in the sample corresponding to the h-th table who are classified as belonging to the i-th sub-population, $N_{h.j} = \sum_{i=1}^{s} n_{hij}$ denotes the marginal total number of subjects in the sample corresponding to the h-th table who are classified as belonging to the j-th response category, and $N_{h..} = \sum_{i=1}^{s}\sum_{j=1}^{r} n_{hij}$ denotes the overall marginal total number of subjects in the sample corresponding to the h-th table. All of these quantities are assumed to be fixed constants rather than random variables. The types of situations where this type of assumption applies are

a. Observational and/or historical data from restricted populations as obtained in retrospective studies, case-control studies, etc.
b. Experimental design data from a strict randomization model point of view;
c. Product multinomial model sample data as described in Appendix 2 from a conditional distribution point of view.

The basic hypothesis of interest for this situation is

H_0: For each of the tables h = 1, 2, ..., q the response variable is distributed at random with respect to the sub-populations; i.e., the data in the respective rows of the h-th table can be regarded as a successive set of simple random samples of sizes $\{N_{hi.}\}$ from a fixed population corresponding to the marginal total distribution of the response variable $\{N_{h.j}\}$.

Under the hypothesis H_0, the following probability model characterizes the distribution of the $\{n_{hij}\}$.

$$\Pr(\{n_{hij}\}|H_0) = \prod_{h=1}^{q} \frac{\prod_{i=1}^{s} N_{hi.}! \prod_{j=1}^{r} N_{h.j}!}{N_{h..}! \prod_{i=1}^{s}\prod_{j=1}^{r} n_{hij}!}$$

From the structure of this model, it follows that

$$m_{hij} \equiv E\{n_{hij}|H_0\} = N_{hi.} N_{h.j}/N_{h..}$$

$$v_{h,ij,i'j'} = \mathrm{Cov}\{n_{hij}, n_{hi'j'}|H_0\} =$$

$$\frac{N_{hi.} N_{h.j} (\delta_{ii'} N_{h..} - N_{hi'.})(\delta_{jj'} N_{h..} - N_{h.j'})}{N_{h..}^2 (N_{h..} - 1)}$$

where $\delta_{ii'} = \begin{cases} 1 \text{ if } i = i' \\ 0 \text{ if } i \neq i' \end{cases}$ and

$\delta_{jj'}$ is similarly defined.

Let $\underset{\sim}{n}_h$ denote the vector of observed frequencies $\{n_{hij}\}$. Let $\underset{\sim}{m}_h$ denote the vector of hypothesis based expected frequencies $\{m_{hij}\}$, and let $\underset{\sim}{V}_h$ denote the hypothesis based covariance matrix $\{v_{h,ij,i'j'}\}$. Let $\underset{\sim}{A}$ be an [(r-1)(s-1) x rs] matrix which is rank independent of within table response sum and sub-population sum vectors (e.g., $\underset{\sim}{A}$ is the Kronecker product of any response contrast basis with any sub-population basis). Then, it follows that an appropriate test statistic for H_0 in a total sense is

$$Q_T = \sum_{h=1}^{q} \underset{\sim}{d}_h' \underset{\sim}{A}' \{\underset{\sim}{A} \underset{\sim}{V}_h \underset{\sim}{A}'\}^{-1} \underset{\sim}{A} \underset{\sim}{d}_h$$

$$= \sum_{h=1}^{q}\sum_{i=1}^{s}\sum_{j=1}^{r} \left(\frac{N_{h..} - 1}{N_{h..}}\right)\left\{\frac{(n_{hij} - m_{hij})^2}{m_{hij}}\right\}$$

$$= \sum_{h=1}^{q} \left(\frac{N_{h..} - 1}{N_{h..}}\right) Q_{P,h}$$

where $\underset{\sim}{d}_h = (\underset{\sim}{n}_h - \underset{\sim}{m}_h)$ and $Q_{P,h}$ is the Pearson Chi-Square statistic for the h-th table. Under H_0, $Q_{P,h}$ asymptotically has the chi-square distribution with D.F. = (r-1)(s-1). Thus, if all the $\{N_{h..}\}$ are sufficiently large, both Q_T and $Q_{TP} = \sum_{h=1}^{q} Q_{P,h}$ have approximate chi-square distributions with D.F. = q(r-1)(s-1).

On the other hand, if many of the $\{N_{h..}\}$ are small even though the overall sample size $N = \sum_{h=1}^{q} N_{h..}$ is large, then Q_T (and Q_{TP}) are no longer appropriate for testing H_0. In this situation, the Cochran-Mantel-Haenszel type of statistic can be used. These have the form

$$Q_{CMH} = \underset{\sim}{d}_{\cdot} \, \underset{\sim}{V}_{\cdot}^{-1} \, \underset{\sim}{d}_{\cdot}$$

where $\underset{\sim}{d}_{\cdot} = \sum_{h=1}^{q} \underset{\sim}{A}_h \underset{\sim}{d}_h$ and $\underset{\sim}{V}_{\cdot} = \sum_{h=1}^{q} \underset{\sim}{A}_h \underset{\sim}{V}_h \underset{\sim}{A}_h'$. Under H_0, Q_{CMH} has approximately the chi-square distribution with D.F. = (r-1)(s-1). Otherwise, it can be noted that Q_{CMH} is directed at average partial association alternatives in the sense that if certain elements of d_h are consistently positive (or negative) across the tables h = 1, 2, ..., q, then these quantities reinforce one another when combined to form $d_{\cdot}$. Also, the fact that significance of Q_{CMH} is evaluated relative to D.F. = (r-1)(s-1) rather than q(r-1)(s-1) represents another aspect of this approach that potentially permits gains in statistical power here.

In some applications, the response categories may be ordinally scaled, in which case location shifts with respect to this scaling often represent the primary types of alternatives of interest. Thus, it becomes advantageous to target the statistics Q_T and Q_{CMH} on certain mean score functions of the type $F_{hi} = \sum_{j=1}^{r} a_{hj} n_{hij}$ where the $\{a_{hj}\}$ represent a reasonable set of numerical values which have been assigned to the set of ordinally scaled response categories. For this purpose, the basic formulas given for Q_T and Q_{CMH} remain essentially the same except that the matrix $\underset{\sim}{A}$ is allowed to vary across tables in the form $\underset{\sim}{A}_h$ and each $\underset{\sim}{A}_h$ is an [(s-1) × rs] basis of sub-population contrast space with respect to the specific linear combination of response categories that pertain to the functions $\{F_{hi}\}$ within that table. In view of the reduced dimension of $\underset{\sim}{A}$ which these modifications involve, Q_T has asymptotically the chi-square distribution with D.F.=q(s-1) and Q_{CMH} has asymptotically the chi-square distribution with D.F.=(s-1).

Finally, if both the response categories and the sub-population categories are ordinally scaled, then certain types of correlation alternatives are often of primary interest. In these situations, it is advantageous to target Q_T and Q_{CMH} on a single function of the type $F_h = \sum_{i=1}^{s} \sum_{j=1}^{r} c_{hi} a_{hj} n_{hij}$ for each table where the $\{c_{hi}\}$ represent a reasonable set of numerical values which have been assigned to the ordinally scaled sub-population categories. Otherwise, the formulas for Q_T and Q_{CMH} remain essentially the same as originally given, except that $\underset{\sim}{A}$ is allowed to vary across tables in the form $\underset{\sim}{A}_h$, each $\underset{\sim}{A}_h$ has only a single row whose elements are the respective products $\{c_{hi} a_{hj}\}$, and the asymptotic chi-square distributions for Q_T and Q_{CMH} have D.F.=q and D.F.=1 respectively.

For further discussion of Model 0 and the various types of statistics which are of interest with respect to it, see Landis et al. [1977].

Appendix 2: Model 0

For the same general framework described in Appendix 1, the $\{n_{hij}\}$ are assumed to follow the product multinomial distribution.

$$\Pr(\{n_{hij}\}) = \prod_{h=1}^{q} \prod_{i=1}^{s} \prod_{j=1}^{r} \frac{N_{hi.}! \, \pi_{hij}^{n_{hij}}}{n_{hij}!}$$

where π_{hij} represents the probability that a randomly selected subject from the (hi)-th sub-population is classified in the j-th response category. The type of situations where this type of assumption is appropriate are

i. Stratified simple random sampling from an infinite super-population where the strata correspond to the qs cells of the h vs i cross-classification;
ii. Simple random sampling from an infinite super-population from a conditional distribution point of view, in which case h vs i is a domain cross-classification;
iii. Certain mixtures of (i) and (ii) where h accounts for the variables in terms of which the stratification cross-classification is defined and i accounts for the variables in terms of which the domain cross-classification is defined.

Let $p_{hij} = (n_{hij}/N_{hi.})$ denote the proportion of subjects in the sample from the (hi)-th sub-population that are classified in the j-th response category. The $\{p_{hij}\}$ represent unrestricted maximum likelihood estimates of the $\{\pi_{hij}\}$. Let $\underset{\sim}{p}$ denote the vector of $\{p_{hij}\}$ and let $\underset{\sim}{\pi}$ denote the vector of $\{\pi_{hij}\}$.

Depending on the nature of the situation under consideration, certain aspects of the response distribution within each sub-population and/or its relationship to the sub-populations can be formulated in terms of functional transformations $\underset{\sim}{F}(\underset{\sim}{\pi})$, and the extent to which there is variation among such functions can be characterized by linear regression models of the type $\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}$. Thus, the principal objectives of statistical analysis include the estimation of the model parameters $\underset{\sim}{\beta}$ and the corresponding predicted values they imply for $\underset{\sim}{F}(\underset{\sim}{\pi})$, statistical tests for hypotheses involving $\underset{\sim}{\beta}$, and statistical tests for the goodness of fit of the model $\underset{\sim}{X}$. For this purpose, two general approaches which have wide applicability to many specific problems of this type are maximum likelihood methods as discussed in Bishop, Fienberg, and Holland [1975] and weighted least squares asymptotic regression as discussed in Grizzle, Starmer, and Koch [1969], and Koch et al. [1977].

Appendix 3: Model 2

For many types of research investigations, data are obtained via probability random samples with complex designs. Some strategies for their analysis relative to the sampled population are discussed in Koch et al. [1975]. However, super-population issues are philosophically more difficult here because the nature of the hypothetical selection process is not necessarily well-defined. For this reason, one simplistic approach is to adopt a Model 0 or Model 1 point of view for this situation.

DISCUSSION

Denis F. Johnston, Office of Management and Budget

Professor Koch's able presentation addresses three fundamental (one might say eternal) questions encountered by statistical practitioners in all fields of application. First is the distinction between the study population and the target population. To provide a crude but common example, what are the implications of relying on statistics for persons of black and other races (excluding whites) as the only available substitute for statistics on the black population? Second, there is the distinction between the variables under study and the concepts they are operationally assumed to represent. To pursue the preceding example, if our interest is in the relationship between education and income among blacks, what are the implications of utilizing data on "years of school completed" as a proxy variable for education, "median personal or family income per year" as a proxy for income and a study population comprising perhaps 90 percent black persons and 10 percent persons of wide but indeterminate ethnic or racial heterogeneity in place of our "target" population of blacks? Third, Koch addresses the role of technical assumptions pertaining to the research design, existing state of knowledge and the statistical objectives to which a particular research design is fitted. A basic question here is the extent to which the underlying assumptions and data requirements of a given research procedure are in fact satisfied by the data available.

Koch recognizes a common theme in these three questions -- the need for a contextual perspective for evaluating the validity of the use of a particular statistical method by examining the specific nature of its given applications in relation to the interpretation of the results obtain-ed in that application. What this seems to mean is that no statistical method is equally valid in all situations or contexts in which it may be applied mechanically. This interpretation is supported by Koch's argument that the proper application of any statistical methodology to practical problems demands a critical re-examination of the research design and the underlying model at each stage of the research process, so as to incorporate the "feedback" information that is yielded by each stage.

In the several papers he has drawn upon in his presentation, Koch offers some useful guidelines to the statistical practitioner for obtaining the optimal amount of information to meet given research objectives under given constraints of time and resources. He provides illustrations of alternative research strategies for obtaining limited information on a given subject at reduced cost and for obtaining more detailed information from the same body of data but at greater cost. In these examples, Koch stresses the importance of retaining a clear understanding of the research objective -- not only what is the problem or the hypothesis being tested, but how much information is required to satisfy that objective at minimum?

The common 'theme' linking these three questions can perhaps be expressed in plainer English: how poor or imperfect can statistics be before they fail to provide any useful information? As Koch recognizes, any attempt to answer such a broad question must be strongly contextual; the illustrative examples he provides only begin to illuminate the enormous range and diversity of statistical applications and the real-world situations wherein these applications are made. In the face of this contextual diversity, any general advice is bound to be of the sort attributed to the Delphic oracles -- e.g., "Collect all the data you can and use good judgment" -- equivalent to the successful stock investor's advice, "Buy low and sell high!"

It is evident that the need for an "interface" between statistical methodology and statistical practice arises out of the imperfect correspondence between statistics as bodies of data drawn from the real world and statistics as a set of methodological principles derived from probability theory and related mathematical concepts. Koch's contribution properly addresses precisely that "interface." But in doing so, he fails to consider a number of constraints that commonly operate in the context of the practitioner's work. First are the resource and time constraints. Nobody ever has, or ever will measure everything that is ideally required; conclusions must invariably be reached on the basis of incomplete or imperfect information. The methodologist can offer useful guidelines for obtaining the minimum information required with maximum efficiency, as Koch does, but he or she cannot provide general guidelines as to how much information is needed or what precision of measurement is required. These issues must be decided by the practitioner in consultation with the client. Second are constraints on communication. If some (many?, too many?) practitioners are less sophisticated statistically than methodological experts, their clients may often be far less sophisticated than the practitioners. To use current jargon, the practitioner must "interface" with a variety of clients whose familiarity with statistical language and concepts is rudimentary at best. This implies that the practitioner must deal with a double problem of translation -- he or she must first adapt the methodologists' guiding principles to the particular context and must then convert the research findings into language that can be understood by the client. This second "interface," between practitioner and client, is at least as important as that between practitioner and methodologist, since it alone assures that statistical findings can be allowed to play a role in public and private policy decisions.

A third set of constraints relates to the decision process itself. The classic portrayal of the statistical practitioner at work is closely similar to that of the practicing scientist: the problem is given by the client and the use

made of the findings obtained is likewise up to the client. Between these limits, the practitioner is expected to utilize the most appropriate techniques within the context of "value-free" principles of objectivity. But for some practitioners, the above delineation of roles often breaks down. The client may have a problem, but the problem may turn out to be different from the one originally expressed. For decisionmakers in particular, a common problem is that a decision has already been reached and the statistical practitioner is expected to provide a veneer of "objective" validation for that decision. Such cases obviously involve basic ethical principles; statistical practitioners cannot legitimately serve as advocates for particular positions unless these positions are supported by objective statistical evidence. But between the ideal of the objective researcher and the outright demand for a hired statistical gun, there is a vast gray area wherein the practitioner must redefine a problem, adjust its requirements to meet the limitations of the available data and resources, and interpret the research findings in order to best serve the client's needs. To be effective in this latter task, the practitioner must try to see the world as the client sees it; yet in doing so, he or she must carefully avoid seeing the data as the client would presumably like to see them. Few methodologists can offer useful counsel in dealing with this kind of communications problem.

Finally, there are the innumerable situational constraints to which Koch makes occasional reference. Here again, the methodologist can only illustrate by a few well-chosen examples the enormous range of phenomena to which statistics find application and the great diversity of circumstances affecting particular applications. By situational constraints we mean the need to recognize and consider the changing social, cultural and historical context from which our statistical observations are obtained. This contextual meaning is insignificant in the many fields of application so favored by the methodological experts -- grain fields, mice in laboratories, and the like. But it is highly significant in the realm of socioeconomic applications, where each statistical observation is subject, in principle, to an interpretation that reflects an historically unique context. A familiar example may suffice to illustrate this point: the rate of unemployment in country A may be strictly comparable with that in country B insofar as both measures employ the same concepts and measurement procedures. But its interpretation may be quite different because of differences in the historical meaning and experience of unemployment in the two countries. The same problem may arise in interpreting identical measures of unemployment in the same country at two widely separate points in time. It is arguable that such interpretations move us far beyond the legitimate purview of the statistical practitioner, but to admit this is to seriously restrict the role of the statistician in addressing complex social problems.

We cannot all be statisticians, and the statisticians among us cannot all possess equal abilities. Hence the "interfaces" between methodological experts and practitioners, and between practitioners and clients are likely to persist as major problem-areas. Koch offers some useful and well-illustrated guidelines for coping with the interface between methodologist and practitioner. Perhaps only the practitioner can develop corresponding guidelines for dealing with the more demanding "interface" between practitioner and the ultimate user of statistical information.

DISCUSSION

Wayne A. Fuller, Iowa State University

Dr. Koch has discussed topics that have long been of concern to statisticians. One of these, the idea of a target population was addressed by survey statisticians in the 1930's and 40's when random sampling of finite populations was being introduced. More recently discussions of "analytic surveys" again brought the topic to the surface. Most sampling texts contain some discussion of target population. On the basis of these discussions one might identify three possible objectives for the estimates constructed from a sample of a finite population.

The first would be: Estimation of a property (a parameter) of the particular finite population sampled. The parameter might be the mean, the difference between the means of two groups, or a regression coefficient. This type of inference problem is, perhaps, most natural and comfortable for the traditional survey sampler. It is the task of a number of government agencies such as the Census Bureau and the Bureau of Labor Statistics.

The second problem is the estimation of a parameter of a finite population separated by time or space from the finite population actually sampled. For example, a study of recreation activities was conducted in Iowa to predict future demand for recreational facilities. This material was requested by the State Conservation Commission as a guide for parkland acquisition, etc.

The third problem is the estimation of a parameter of an infinite population from which the finite population is a conceptual random sample. I think most will agree that scientists are often interested in inferences beyond the finite population studied. This does not mean that it is always easy to define the conceptual population of interest.

One might place the three objectives in a hierarchy, the estimation of the particular finite population parameter being the narrowest objective and the estimation of the infinite population parameter the broadest. However, a careful consideration of the problem of estimating for a second finite population seems to require a specification of the relationship between two finite populations. This in turn leads one to the infinite population concept.

When only one population is sampled it seems that the statistician can only help the subject matter specialist assemble and interpret data on which to make the judgment on comparability. On the other hand, if we have sampled a number of finite populations, for example, a number of years, we may be able to bring statistical analysis to bear on the nature of the comparability of the finite population of interest (next year). That is, one might formalize that problem by assuming that the sequence of finite populations was a realization from a common generating mechanism.

Let us consider briefly the idea of a superpopulation. One does not have to be an authority on the history of statistics or on the foundations of statistics to recognize that the ideas of superpopulation permeate the literature. For example, Fisher (1925, p. 700) in a prefatory note to his 1925 paper "Theory of Statistical Estimation" stated, "The idea of an infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability." Also, little reading is required to establish the diversity of opinions statisticians hold with respect to the ideas of superpopulation. An idea of this diversity can be obtained by reading the volumes New Developments in Survey Sampling edited by Johnson and Smith (1969) and Foundations of Statistical Inference edited by Godambe and Sprott (1971).

In many of the studies of sample survey data falling within our personal experience, the investigator was interested in conclusions beyond the finite population actually sampled. As I said before, this does not mean that the investigator could perfectly specify the population of interest. If the statistician poses the question, "For what population do you wish answers?" he should be content with a rather vague answer. In fact, the answer "I desire inferences as broad as possible" will be a reasonable reply in the minds of many scientists. Such an answer means that the investigator wishes a model with the potential for generalization. Given this desire, the statistician should assist in constructing models with that potential.

Treating the finite population as a sample from an infinite population is one framework which provides the potential for generalization. In fact, I believe a strong case can be made for the following position: "The objective of an analytic study of survey data is the construction and estimation of a model such that the sample data are consistent with the hypothesis that the data are a random sample from an infinite population wherein the model holds." While this statement is something of an inversion of the manner in which the traditional statistical problem is posed, it seems to be consistent with the manner in which scientific progress is made.[1]

When presented with analytic survey data I believe one constructs models acting as if the data were a sample from an infinite population. (Of course one should not ignore the correlation structure of the sample data. Correlation among sample elements may arise from properties of the population or may be induced by the sample design. For example, if the sample is an area sample of clusters of households, the correlation between units in the same area cluster must be recognized in the analysis.)

A scientific investigator reports carefully the procedures, motivations, and alternative postulated models associated with the analysis. Those things considered unique in the material

(the nature of the sample) are reported together with the findings for that material. The reader of the scientific report must decide if the results of the study are applicable to the reader's own problem.

Let me give a preface to my next remarks. When the originally scheduled third discussant was unavailable, it was decided to replace him with a biometrician, in order to add balance to the group of discussants. Time was short and biometricians were in even shorter supply. I was tapped for the position by a biometrician who is not attending the meetings. Hence, I feel a certain obligation to biometricians in general, if not to the absent member of that group.

Therefore, in my role as a biometrician, I would like to emphasize the importance of the knowledge of "biology" (or other subject matter fields) in model construction. Let me do this with an illustration. I have never used stepwise procedures in constructing models for empirical data. I have always felt that the subject matter person and I should actually specify an array of possible models at every step of the process. I feel that we should be better able to specify a model than a machine. This does not mean that we do not try alternative models or that we are blind to the data. Preliminary summaries, plots, and residual analyses are used. But I feel that it is important to think about the material using all available knowledge, intuition, and common sense at every step of the model building process. It seems to me that real effort is often required to persuade a subject matter person to share his knowledge with his statistical consultant. Perhaps it is because his knowledge is vague, based on analogy and conjecture. But it is precisely the kind of knowledge that should be fed into the model building process. Working together in specifying models often brings this kind of information to the surface. As Leslie Kish said last night, statisticians and statistical methods are powerful tools available to the scientist. They are not substitutes. The really successful consultant never forgets this fact. The first question, the last question, and the question at all steps between is: Does it make sense?

Dr. Koch mentioned that the variables we observe are often imperfect representations of the concepts that interest us. There are at least two levels to the problem. The first level is the failure to obtain the same value for a particular variable in different attempts to measure it. This kind of error is called response error in survey methodology and measurement area in the physical and biological sciences. If the independent variable in a simple regression is measured with error, the coefficient is biased towards zero. In the multiple independent variable case, the effects of measurement error are pervasive, but not easily described. If the error variances are known (or estimated from independent sources) there are techniques available for introducing that knowledge into the estimation procedure. I feel that this is an area that deserves more emphasis in the "statistical methods" literature.

The second level of the problem is more subtle. Consider an IQ test. The repeatability of such tests is fairly well established and the reliability (a measure of the relative error variance) is often published with the test. Yet we realize that the mean of an individual's test scores is not perfectly correlated with that illusive concept we call intelligence. It may not even be linearly related (the scale problem). Thus, we must always be on guard against drawing incorrect conclusions by treating a variable as if it is perfectly (or even linearly) related to our concept. My colleague, Leroy Wolins, has collected a file of applied papers that he believes contain errors of the second kind.

I close, believing that the items we have been discussing will be of concern to statisticians and scientists for years to come.

FOOTNOTES

1/ I believe that Kempthorne and Folks (1971, p. 507) come to this position in their discussion of Pierce.

REFERENCES

[1] Cochran, W. G. (1946), Relative accuracy of systematic and stratified random samples for a certain class of populations. Ann. Math. Statist. 17, 164-177.

[2] Cochran, W. G. (1963), Sampling Techniques. Wiley, New York.

[3] Deming, W. E. (1950), Some Theory of Sampling. Wiley, New York.

[4] Deming, W. E. and Stephan, F. F. (1941), On the interpretation of censuses as samples. J. Amer. Statist. Assoc. 36, 45-59.

[5] Fisher, R. A. (1925), Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22, 700-725.

[6] Fisher, R. A. (1928), Book review, Nature 156-196.

[7] Godambe, V. P. and Sprott, D. A. (1971), Foundations of Statistical Inference. Holt Rinehart and Winston, Tronto.

[8] Johnson, N. L. and Smith, H. (1969), New Developments in Survey Sampling. Wiley, New York.

[9] Kempthorne, O. and Folks, L. (1971), Probability, Statistics, and Data Analysis. Iowa State University Press, Ames, Iowa.

[10] Madow, W. G. (1948), On the limiting distribution of estimates based on samples from finite universes. Ann. Math. Statist. 19, 535-545.

PRIVACY AND CONFIDENTIALITY AS FACTORS IN SURVEY RESPONSE

Edwin D. Goldfield, National Academy of Sciences; Anthony G. Turner and Charles D. Cowan, Bureau of the Census; John C. Scott, University of Michigan

In recent years there has been much discussion among survey practitioners about perceived growing difficulties in conducting surveys of human populations. In 1973, under a grant from the National Science Foundation, the American Statistical Association brought together a group of social scientists and survey methodologists to explore the problems and to try to determine whether they constituted a threat to the continued use of surveys as a basic tool of social science research. The conference, meeting in May and December, reached five general conclusions [1]: (1) That survey research is in some difficulty; (2) to an undetermined scale that difficulty is increasing; (3) the problem varies in incidence between government, private and academic research; (4) the grounds for concern are great enough to urge the prompt initiation of a more intensive examination of the problem and programs to meet it; and, (5) there are many potential areas for action, some of which could start now.

In Lester Frankel's presidential address to the ASA in 1975 [5], he discussed the problems of maintaining satisfactory response levels in surveys, and gave attention to the public's fears of invasion of privacy and violation of confidentiality of records as a contributing factor. Marketers, political scientists, and other producers and users of survey data have also been actively concerned [2,6,8,10]. Newspaper writers have reported back to the public the concern of survey takers and users about public reaction to surveys [11].

While a number of reasons have been adduced for the reported increasing difficulties in obtaining information through surveys--fear of crime, changes in living and working situations, oversurveying, disillusionment about the validity of survey results, salesmen masquerading as survey takers--concerns about privacy and confidentiality receive prominent mention as a cause. There seems to be general agreement that there is an insufficiency of empirical, quantitative information on current trends in response rates[1] (or in the level of effort needed to maintain response rates) and on the factors that may be associated with changes. One of the putative factors that is especially difficult to quantify is that of privacy and confidentiality concerns.

The Bureau of the Census has undertaken to try to discover what the feelings of the public are and how they affect the public's behavior as respondents in censuses and surveys. As part of this effort, it commissioned the Committee on National Statistics of the National Academy of Sciences to participate with it in an exploratory study. The Committee established a multidisciplinary group of experts, the Panel on Privacy and Confidentiality as Factors in Survey Response, chaired by former ASA president William H. Shaw. The Panel has outlined a number of avenues of investigation, and has participated with the Bureau of the Census and with the Survey Research Center of the Institute for Social Research at the University of Michigan in carrying them out. These investigations are in progress; this paper will describe them, with particular emphasis on the two surveys that are major parts of the overall study[2].

The Panel recommended that two fairly small-scale exploratory surveys be taken to test the feasibility of obtaining some quantitative evidence on people's opinions and behavior with respect to surveys. One of them is a survey of recalled past experience as survey respondents (or nonrespondents) and of attitudes about surveys, conducted jointly by the Bureau of the Census and the Michigan Survey Research Center. It is recognized that attitude surveys may not be reliable predictors of behavior. However, it was felt that the kind of attitude survey that was tested might indicate its value in blocking out areas of concern or nonconcern and areas of knowledge or ignorance, and might indicate differences between population groups.

The second kind of exploratory survey that the Panel recommended is of a different nature. It is an experiment in measuring response behavior, in particular, differential response behavior when confronted with promises of confidentiality differing in duration of protection. The legal conditions under which the Census Bureau operates cause it to be especially interested in this aspect, although other data-collecting and data-holding organizations can also be expected to be interested. The Census law (Title 13, U.S. Code) requires the Bureau to keep confidential, even from other Federal agencies, the individually identifiable information it collects. However, there is one ambiguous dimension to the assurance of confidentiality, and that is its duration. The Census law does not specifically state whether the confidential status of the individual data is to endure forever or for some limited period of time. A law pertaining to the National Archives of the United States suggests that confidential government records are not to be kept under lock and key forever. Under an agreement pursuant to that law, the 1900 census records, in the custody of the National Archives and Records Service, have been opened to researchers, and it is the intention of the Archives to open each succeeding set of census records as it reaches 72 years of age. There is much advocacy by researchers for still earlier access to census records, e.g., after 50 years, or even 10 years. Bills have been introduced in the Congress to specify one period of confidentiality or another. The Census Bureau, which has been accustomed to promising confidentiality without an end date, is concerned about whether it can expect good public cooperation in the 1980 census if its

Walt R. Simmons, NAS, has made major contributions to the planning and conduct of this study.

confidentiality promise for that census is equivocal or if it specifies a limited period. It has had no real evidence on what is or is not acceptable to the public. The surveys are designed to cast some light on this question; they are described in some detail in the later portions of this paper.

In addition to the two surveys, the project has been exploring some other avenues. It was recommended that opportunities be sought to conduct semi-structured discussions about privacy and confidentiality with selected small groups. It was felt that interplay within the group might bring out and develop ideas and feelings more clearly than could be done by other means such as individual questionnaires. A number of such small-group discussions have been held, by the Census Bureau and by the Survey Research Center. They have provided a good deal of interesting material for analysis as a separate part of the study, and also were useful in planning the questionnaire content for the attitude survey. These sessions involved, in separate groups, Census Bureau interviewers, Survey Research Center interviewers, SRC staff, members of ethnic and church groups, members of a women's civic organization, and senior citizens. While it is difficult and hazardous to generalize from such experiences, some impressionistic findings suggest themselves. Participants (other than the survey takers themselves, and even some of them had doubts) tended to concur almost unanimously in a disbelief in the confidentiality of individual records. (Findings of the attitude survey were consistent with this expression of skepticism.) Different subjects of inquiry were regarded as having quite different degrees of sensitivity. Income was commonly mentioned as an objectionable topic. Others included sexual behavior, number of children expected, marital discord, and inquiries about neighbors. People saw little concrete evidence of the value of surveys; they said they would be more willing to participate in a survey if the benefits were explained beforehand. People had negative feelings about surveys not only because of their perceived invasion of privacy, lack of confidentiality, and failure to yield tangible benefits, but also because the survey approach was thought of as often employed as a sales or crime ruse. Despite these adverse views, there were indications that people would be willing to cooperate in surveys if approached in a convincing and reassuring manner; it seems clear, however, that this is not easily accomplished.

Another phase of the project is a review of relevant literature and a canvass of selected survey research organizations, both governmental and nongovernmental. A majority of the approximately 30 survey organizations that replied to the inquiry reported that current response rates are lower than they were five to ten years ago, or that it now requires more effort to secure the same level of response. Increases in refusal rates were reported, along with increased difficulties in contacting designated respondents. A mitigating circumstance was the improvements reported by some survey organizations in their sample designs and survey procedures. These changes may have a positive effect on the quality of survey results counteracting the negative effect of increased difficulties in respondent contacts.

Research Design of the Behavioral Experiment

In order to test the effects on response of varying promises of confidentiality, a designed experiment was developed and carried out. The research design was a classical application of controlled experimentation in the field of human surveys, utilizing randomized blocks. A nationwide multistage probability sample of 502 clusters of 5 households each was selected in 20 Primary Sampling Units (PSUs). Within each cluster, households were assigned randomly to one of 5 treatment groups and personal interviews were conducted by Census Bureau interviewers during September 1976. Interviewer assignments consisted of whole clusters so that each interviewer administered all 5 treatments for a given assignment. The survey was voluntary. The content of the questionnaire was identical in all 5 treatments and consisted of items comparable to those which appear in a decennial census, including population variables such as sex, age, marital status, educational attainment, and income; and housing variables such as tenure, plumbing facilities, value of property, and rent. Only the interviewer's introduction, which was read verbatim to the respondent, was varied as follows: Treatment A--"Your home is among those selected for a nationwide survey being conducted by the United States Bureau of the Census. The survey is authorized by title 13, United States Code; participation in the survey is voluntary, and there are no penalties for refusing to answer any question. However, your cooperation is extremely important to insure the completeness and accuracy of the final results. This survey collects basic information about population and housing, and will help to prepare for the Twentieth Decennial Census which will be taken in 1980. Your answers to this survey will be used only to form statistical totals and averages that will not identify you personally in any way. Your answers are confidential and will never, at any time, be given to any other agency or to the public." Treatment B-Same as A, except the final sentence is "your answers will be kept confidential for 75 years; however, after that time they may be given to other agencies and to the public." Treatment C--Same as B, except the duration is 25 years. Treatment D--Same as A, except the final two sentences are deleted. Treatment E--Same as A, except the final two sentences are "your answers will be used to form statistical totals and averages. Your individual answers may also be given to other agencies and to the public."

The first-stage selection units, as mentioned, were 20 PSUs chosen throughout the U.S. The second stage of selection consisted of 502 clusters, or segments of housing units. These were noncompact clusters with an expected size of 20 units each. For a randomized block design such as this one, it would have been better to select compact clusters of 5 units each to maximize the homogeneity within each block (cluster). However,

there was no way of insuring in advance that the 5 compact units selected would all be eligible for interview. Since it was critical that for each cluster all 5 (and only 5) treatments be administered, it was desirable to minimize the chances of discarding clusters because they contained vacant or demolished units or others ineligible for the experiment. Therefore it was decided to select 20 noncompact units, have the interviewer canvass them for eligibility, and systematically select 5 of the ones determined to be eligible. Units were determined to be eligible in the precanvass (which involved personal contact where necessary) if they were currently occupied and the residents were not away on vacation or other extended absence.

Also influencing the decision to use noncompact clusters was the need to lessen the possibility of a potential bias in the administration of the survey. Because of the Census Bureau's law (Title 13) governing the confidentiality of data, it was decided the respondents in this research project would ultimately have to be told that the answers they supplied would be confidential forever, irrespective of the particular stated condition of confidentiality given them prior to the interview. A letter, therefore, was left behind with each respondent following the interview. The letters varied somewhat depending upon treatment type, but they essentially explained the nature of the experiment and informed the respondent that the answers were indeed confidential forever in accordance with present law, in spite of what was said at the outset. Because the letter, in effect, let the cat out of the bag, there was concern about possible biases if close-by neighbors were to discuss the experiment when one of them was scheduled to be but had not yet been interviewed. It was expected that using noncompact clusters would reduce the chance of bias of this type from occurring.

The third stage of selection in this experiment involved choosing exactly 5 of the units determined to be eligible out of the original expected 20. This selection was done by the interviewer through the use of a random selection table. Moreover the order in which the 5 selected units was assigned to treatments was also randomized, so that the geographic ordering of the 5 selected units was not always in the same pattern, such as ABCDE. It was felt that this procedure was necessary to inhibit interviewers from, subconsciously perhaps, arranging the sample units in some biased fashion. Another important feature of the sample design was the stratification employed for oversampling nonwhite households. The anticipated overall sample size of 500 households per treatment was too small to detect treatment differences among important subgroups of the population. Therefore, clusters containing a high proportion of nonwhite households were selected with a probability double that of the remaining households. The criterion for stratification was that Census enumeration districts (ED's) which contained 20% or greater nonwhite households in 1970 made up stratum 2 while all remaining ED's made up stratum 1; new construction units, for which there was no a priori information on racial composition, were included in stratum 1.

In choosing the specific treatments to be tested, several considerations were taken into account. Treatment A households constituted the control group inasmuch as they were given the standard Census Bureau promise of confidentiality. The choice of a 75-year promise of confidentiality as one of the treatments (B) actually represented a very real practical possibility since legislation has been proposed to make confidential Decennial Census records available to historians and other researchers through National Archives access after that period of time. It was important also to use a treatment group that might reflect a more meaningful impact on respondents while they were still alive rather than strictly upon their descendants, since very few adult respondents would be living 75 years hence. Twenty-five years was therefore chosen as a third treatment group (C). The choice of no confidentiality at all was an obvious one, but it was felt that an important distinction in the research design would have to be made between an explicit statement of no confidentiality and an implicit one. One of the objectives of the total program was to ascertain the degree to which confidentiality concerns contribute to survey nonresponse. It was not known a priori whether confidentiality of information is a dominant factor in a respondent's mind when he agrees or does not agree to participate in a survey. As a result, treatments D and E were both used, with D giving no confidentiality by inference and E explicitly stating nonconfidentiality.

There was much concern as to whether our interviewers could carry out this project unbiasedly. Census interviewers have been trained on all other Bureau surveys to know that Census data are confidential. Many of the interviewers use the fact of confidentiality to persuade reluctant respondents to grant an interview. Such behavior could not be tolerated in this experiment. Because of the importance of the survey, it was desirable to use senior-level interviewers rather than newly recruited ones insofar as possible. Presumably, new interviewers would have been less influenced by prior knowledge of Census confidentiality safeguards. The interviewers were given a one-day training session which, among other things, emphasized the nature of the research objectives, and the requirement to avoid mention of data confidentiality in trying to persuade reluctant respondents to participate.

Results of the Designed Experiment

The analysis plan for the designed experiment was to consist of a comparison of refusal rates by treatment; secondly, there was to be an examination of item nonresponse by treatment. Thirdly, the question of differential response validity by treatment was to be addressed if possible. Finally, a series of questions at the end of the interview was to be analyzed to shed some light on how well the respondent paid attention to or remembered the interviewer's opening statement.

With regard to the overall refusal rate, the survey procedures called for the interviewer to record appropriate information about what point in the attempted interview a refusal was actually encountered. It was of key significance in the design objectives to know, for example, whether refusals occurred before or after the interviewer read the introduction. The estimation scheme that was employed was one that preserved the differential probabilities of selection of the sample units but which did not inflate the data to national totals, since no useful purpose could be seen by doing the latter. No adjustment was made for nonresponse, of course, since nonresponse (especially refusals) was the statistic we sought to study. Of the original 502 clusters selected, 14 were eliminated from the survey because fewer than 5 of the expected 20 units in each of these clusters turned out to be eligible for interview. This situation usually occurred because large, sample buildings had been demolished. The final survey thus contained 488 clusters of 5 units each, or 2440 households.

Table 1 shows the nonresponse statistics by treatment for the two strata combined, properly weighted to account for the double probability of selection of stratum 2 households in relation to stratum 1 households.

There is an indication of a possible interviewer effect in the distribution of no-one-home noninterviews. Examination of only those treatments (ABCE) where confidentiality was explicitly mentioned reveals a monotonic increase in the no-one-home noninterviews as the degree of confidentiality decreases. In the course of listing the units for eligibility determination in the sample segments, interviewers often had to inquire at the housing units to obtain current occupancy status. One could conjecture that for households where this initial contact was met with respondent hostility, some interviewers could have acquired the tendency to accept a no-one-home NI more readily if the unit were subsequently sampled and assigned to treatments other than A. The assumption is made that it was easier for the interviewers to approach treatment A households, in spite of exhortations to them in the training to apply equal attention and care to all households in all treatments.

The last line of Table 1 is perhaps the chief result of the entire experiment, for it shows the key refusal rates by treatment for those respondents who were exposed to the treatment variations. For those households where no one was at home or the refusal occurred before the statement was read, the nonresponse should be independent of the statement variation. It is difficult to draw definitive conclusions about the degree of difference by treatment because the observed differences are small and are generally within sampling error.[3] For example, the largest estimated difference between treatments for the key refusal rates, as shown in Table 1, is between Treatment E (2.8%) and Treatment A (1.8%). This difference is estimated at 1.0 percentage point with a standard error of 1.2 percentage points. Hence the observed difference is not significant even at the 68% level of confidence.

Of course, it is not only the magnitude of the treatment differences which is important, but also their pattern. The trend in the key refusal rate of increasing refusal with decreasing assurance of confidentiality suggests other tests for assessing pattern significance. First, however, the heuristic observation can be made that no such trend is present for refusal rates before the statement was read. Such a trend would be indicative of design or execution flaws somewhere. One test which was applied was an attempt to discover the existence of a linear trend in the post-statement refusal rates, the presumption being that the values of the proportion refused should increase as we move from Treatment A to Treatment E. For this purpose scale values must be assigned to the treatments. The values chosen were 3 for Treatment A, 2 for B, 1 for C, 0 for D, and -1 for E. The procedure simply involved testing the null hypothesis that the regression coefficient, b, of p_i on X_i is equal to zero, where p_i and X_i are the proportion refused and the assigned scale value, respectively, for the i-th treatment group. The regression coefficient and its standard error were calculated in accordance with the Snedecor-Cochran [12] procedure, except that weighted values were used to account for the double probability of selection of sample cases in stratum 2. The computed regression coefficient and its standard error were -0.00278 and .00165, respectively (see Table 2). The corresponding t-statistic is -1.69. We would conclude therefore that the trend is statistically significant at the 90% level of confidence.

The test for a linear trend, as carried out in Table 2, has two objections however. First, our data were not chosen in a simple random sample and secondly, the assignment of scale scores (X_i's) is more or less arbitrary. The observed trend can also be examined for significance by using two nonparametric tests which have the advantage of being free from constraining assumptions about the distribution of the population or the nature of the sample design. The first is Spearman's rank correlation coefficient which can be used as a measure of the degree of concordance between the hypothesized and observed ranks of treatment refusal rates. In this experiment it was hypothesized that refusals would increase with decreasing assurance of confidentiality which is precisely what the empirical evidence supports. Table 3 shows that the correlation between the hypothesized and observed rankings is statistically significant with 95% confidence.

Kendall's τ can be similarly employed as a measure of concordance between hypothesized and observed rankings of the treatment refusal rates. This statistic, shown in Table 4 yields statistical significance at approximately the 99% level.

On the whole one could conclude therefore that it is improbable that the observed pattern of refusal rates would occur if in fact the underlying refusal rates were the same for all treatments. With the sample size employed for this study, however, one cannot reliably estimate the magnitudes of the refusal rate differences among treatments.

Aside from the question of the trend in the key refusal rate by treatment, two other observations are noteworthy from Table 1. The first seems to be that irrespective of the stated condition of confidentiality the refusal rates, by nearly any standard, are not large. It is not clear whether this result is due to general lack of concern on the part of the responding public about what happens to information they furnish officialdom or whether there is an undergirding of citizen trust in the Census Bureau insofar as the uses it makes of data it collects. It remains to be seen whether less than total confidentiality affects the validity of response, however. This question will be addressed by a validation study that was undertaken, results of which have not yet been compiled.

The second observation concerns the refusals recorded before the interviewer actually read the statement outlining the confidentiality conditions. Here there is an overall weighted refusal count of 123 which is somewhat higher than the 95 recorded for refusals after the statement. We would interpret this to mean that for a little more than half the people who were inclined to refuse this survey, it appears clear that confidentiality specificity was not the determining factor. This is not to suggest, however, that concern for confidentiality played no role in their decision; it is conceivable that an unknown number of them could have held a priori opinions that this survey (or possibly any other government survey) was not in their best interest vis-a-vis confidentiality safeguards.

Differential analysis for the high nonwhite stratum turned out to be fruitless because the number of key refusals was so small. The raw number of refusals in this sector ranged only from 0 to 2 for a treatment class, and there was only a total of 7 refusals in all of the 5 treatments combined. Also, it was mentioned earlier that item nonresponse was part of the plan for analyzing the treatment effects. It was hypothesized that some respondents might agree to answer some of the survey questions rather than refuse the entire interview outright, but there might be a tendency for individual question refusals to increase as the promise of confidentiality protection decreased. Neither space nor time permits a thorough examination of the data here. In general it can be reported that the sociodemographic items showed very little item nonresponse nor any significant differential by treatment in nonresponse for the item.

It was of methodological interest in this study to determine the relative efficiency of the randomized block design in case a larger scale survey is done. A two-way analysis of variance would have been the appropriate technique for making this determination. There was, however, no computationally convenient method of coping with the complicating problem of missing values due to nonresponse for reasons other than refusal after the confidentiality statement was read; hence the sample size was not constant by treatment. Moreover, the sample was not chosen in a simple random fashion. Some information can be brought to bear on the question of blocking efficiency by considering the covariances among treatments with respect to the target statistic, that is, refusals. In carrying out the computations it was discovered that the covariance estimates made a trivial contribution to the total variance of the estimated difference between any two treatment refusal rates. By inspection the reason for this result can be attributed to the fact that refusals to more than one treatment rarely occurred within the same cluster or segment. In fact there were only 3 segments in the total of 488 that had multiple refusals and all 3 had only 2 refusals. We conclude therefore that the blocking was not particularly beneficial, at least with respect to the key statistic of interest, refusal after a stated confidentiality variant.

The concluding section of the questionnaire contained a few questions to ascertain how well the respondent remembered the opening confidentiality statement by the interviewer. The frequency distributions by treatment for these items are shown in Tables 5 through 9. According to these results the respondents did a respectable job in listening to and recalling what was said about confidentiality. Seventy-four to eighty-two percent (Table 5) of all persons said they remembered that a statement was read, and for Treatments A, B, C and E, 50-77 percent (Table 6) recalled that confidentiality was mentioned. The distribution of persons who responded that confidentiality was promised is in accord with the actual statements made (Table 7). Tables 8 and 9 are good reflections of the facts. There is some suggestion that there is a carryover effect of Census Bureau reputation and/or publicity that leads people to believe the data are confidential, despite what the interviewer may have said. For example, 40 percent of persons with Treatment D said that confidentiality was mentioned with 26 percent claiming it was promised, even though the interviewer had said nothing about the subject. Moreover, 22 percent of the Treatment E group claimed the interviewer gave them a promise of confidentiality when in fact she did the opposite.

Design of the Attitude Survey

The attitude survey was designed to measure the feelings of the public about privacy and confidentiality, and how these factors might affect survey response. The survey tried to measure indirectly reactions to being surveyed by asking about prior survey experience, and whether prior survey contacts were seen as invasions of privacy, or whether prior contacts had led to unpleasant or adverse situations later, even in cases where confidentiality had been promised. The survey continued in its indirect approach by asking questions concerning trust in survey results, survey organizations and government. These questions, combined with some knowledge questions on surveys, provided a backdrop for questions directly related to confidentiality, and in themselves were an index to a respondent's willingness to be interviewed by the government. Direct questions regarding privacy and confidentiality included whether the respondent knew how long

Census records were confidential, how long should the records be kept confidential, and who really had access to the records. Finally, the respondent was given a self-administered form which asked for his reactions to the survey in which he had just participated.

Because of a concern that responses to the government about the government may be tainted by respondent tendency to be accommodating or polite to the interviewer, the decision was made to divide the data collection with the Survey Research Center at the University of Michigan. Dividing the field work allowed testing to see whether auspices had any significant effect on response to questions about the government. The design employed also permitted internal reliability checks between independently managed half-samples. An essential feature of the design was that it was a national probability sample of the coterminous U.S. which was split into two interpenetrating parts. These parts were then randomly assigned to SRC and Census, with each agency conducting interviews in its assigned half-sample. The sample for the study was drawn by the Survey Research Center.

The sample was located in 44 PSUs of SRC's national sample. At the second stage, segments with an expected size of 8 to 16 housing units were chosen within the primary areas. These segments were listed by the SRC interviewers, and an average of 8 housing units per segment designated for interviewing. Every second selected listing from a random start was assigned to subsample A and the remaining selections assigned to subsample B. This procedure yielded approximately 860 listings per subsample. A random assignment of these two subsamples was then made between Census and SRC. At the third stage, within-housing-unit randomized selection tables were used to make a probability selection of one designated person from all residents 18 years of age or older in each of the selected housing units. Thus, while the housing unit selection probabilities were equal within subsamples, the selection rates within housing units varied by the number of eligible adults.

Regarding the development of the questionnaire, a topic outline with draft questions on major topics together with the transcripts from a series of several small group discussions (previously mentioned) served as the basis around which the initial version of the questionnaire was constructed. The questionnaire was extensively revised during two pretests. The pretests showed that direct questions about the isolated concepts of privacy and confidentiality produced reports of high sensitivity and concern, but that respondents were willing to trade off these values to maximize other values when faced with specific situations. It was as if people were saying "yes, we like apple pie" but then passing up a serving because they were on a diet. Because of the problems encountered with questions about abstract concepts, the focus on the final questionnaire was placed on the respondents' direct experience with surveys. The survey instrument itself served as a standard treatment, incorporating many "typical" demographic questions, and reactions to the questionnaire were gathered on a self-administered form presented to the respondent at the end of the interview.

Regular staff interviewers were used by both organizations. These interviewers can be characterized as experienced and mature. They were primarily women with more than a high school education and were typical of those working for the two interviewing organizations. All specific interviewer preparation on this study was done by written instructions developed by SRC but used by both organizations. Written instructions rather than classroom training were used to guarantee standardization of procedures and preparation between organizations.

One slight deviation from normal procedures was that no advance letters were sent to the sample housing units. This was done to avoid the possibility that neighbors might become concerned if one received notification that an SRC interviewer would be calling and the other got a letter from Census. We have no evidence that this study's response suffered from not having an advance letter. The two organizations maintained close communication during the interviewing period to coordinate efforts and assure standardization of procedures. All editing and coding of the interview content was handled by the Survey Research Center to assure processing comparability between the two half-samples. The code books were constructed by SRC in consultation with the Census.

Two follow-up efforts were made in conjunction with the attitude survey. The first was an attempt to obtain information from nonrespondents by mail, and the second was a very small reinterview survey (using the attitude questionnaire) of people contacted on previous studies conducted by Census or SRC. The attempt to learn about nonrespondents by mail failed to produce useful information since only ten people returned the mail forms. The reinterview of people contacted on previous studies was designed as a validation of the survey contact questions contained in the attitude questionnaire. These validation results have not yet been fully analyzed.

Results of the Attitude Survey

Aside from the response rate there was very little difference in the results between the SRC and Census half-samples. Most of the results presented here will therefore be for the combined samples. The overall response rate on this study was 81.9% for both SRC and Census combined. Census achieved a response which was 6.7 percentage points higher than SRC. The difference in response rates between the two organizations was found to be concentrated in large SMSAs, in refusals (as opposed to other types of NI) and in interviewing persons over 65. In each instance Census achieved significantly less nonresponse than SRC.

The range of the questions in the attitude survey allows for a great deal of analysis to be done.

Only a few of the highlights of the survey results can be presented here. The survey tried to tap feelings about the relation between Census records and privacy using different techniques. In the most direct approach, respondents were asked, "Do you happen to know whether these records (the individually identifiable survey records) are public so that anyone who might want to see them can, or are they not open to the public?" followed by, "Do you know whether individually identifiable Census records are available to other government agencies or not?" The third question in the sequence was "Do you feel that other government agencies could obtain individual records from the Bureau of the Census if they tried?" Table 10 shows the combined results for both the SRC and Census half-samples and reveals that 18 percent of the respondents believe that Census records are open to the public, another 22 percent believe that Census records are open to other government agencies, and another 40 percent believe that other government agencies could obtain confidential Census records if they really tried. This last question was asked only of those respondents who had not indicated they believed Census records to be open to the public or to other government agencies. Of the respondents who were asked, therefore, whether they believed the Census Bureau could maintain confidentiality, 2 of 3 respondents did not feel the Census Bureau could. Overall, 80 percent of the respondents did not believe or know that Census records are confidential, or did not believe that confidentiality could be maintained. An additional 15 percentage points of the remaining 20 percent said they did not know whether the Census Bureau can maintain confidentiality, leaving only five percent of the respondents who were willing to commit themselves on the inviolable confidentiality of Census records. When asked, however, how long Census records should be kept confidential, 46 percent, close to half, of the respondents said the records should be confidential forever. Those respondents who stated that the records should be open after a time were asked, "How long after (the records) are gathered should it be before they are available for researchers outside the Census Bureau?" Of those who gave a numeric answer, the average number of years was 19.5 years.

Though only five percent of the population know or believe that Census records are completely confidential, 46 percent believe the records should be confidential forever, and an additional 40 percent believe the records should be confidential for some time. This means that whereas most people desire their records be kept confidential, they do not know that Title 13 protects their records, or they are skeptical of the Census Bureau's ability to carry out its legal duty. Other questions indicate a rather low level of knowledge about Census. When asked whether the Decennial Census is mandatory, 50 percent of the respondents said yes, 25 percent said no, and 25 percent did not know. Another question reveals that only 45 percent of the population know that the national government conducts the decennial census, and only 31 percent know that the Census Bureau conducts it.

This low level of knowledge about Census and safeguards on the confidentiality of Census records indicate a cause of skepticism among respondents about the Bureau's ability or willingness to maintain confidentiality. Another possible contributing factor to this skepticism is a distrust of survey organizations and earlier contacts by survey organizations.

When asked about organizations that run surveys, 52 percent of the respondents said they felt people were more likely to give accurate information to some types of organizations than to others, while 41 percent said there was no difference between organizations. Of those who said there was a difference in accuracy of reporting to organizations, 37 percent of the SRC respondents said the National Government was most likely to get accurate reporting, whereas 42 percent of the Census Bureau respondents chose National Government as most likely to get accurate reporting. Of the SRC respondents, 29 percent said that universities were most likely to get accurate reporting, whereas only 16 percent of Census respondents said the same (see Table 11).

When asked which type of organization was least likely to get accurate reporting, "private companies" were chosen by 60 percent and 54 percent of the respondents for SRC and Census respectively. The National Government was mentioned by 17 percent of the SRC respondents and 15 percent of the Census respondents, whereas for mentions of universities as least likely to get accurate information, the percentages were 4 and 15 for SRC and Census respectively (see Table 12). And when asked how often can you trust the results of surveys, 41 percent of the respondents to both organizations said that surveys can be trusted almost always or most of the time, and 51 percent said that surveys can be trusted only some of the time or hardly ever.

The results suggest there is a general lack of trust in survey results in a large part of the population. It might be conjectured here that since trust in the National Government has been a topic of discussion in recent years, a carryover effect on Census as a branch of the government could be showing up. There is more trust in the government's ability to collect accurate information than in other organizations, even when one takes account of the halo effect due to having the government ask questions about itself. But in general people are skeptical. This skepticism seems to translate directly into disbelief when asked about confidentiality. If the public is concerned about the trust it places in the government and in surveys, it would hardly trust or believe in the safeguards associated with surveys that the Census Bureau offers. There is a belief by the general populace that Census records should be kept confidential, but there is little knowledge of or trust in the Census Bureau and its ability to maintain confidentiality.

Why do respondents answer surveys then? In the small group discussions when this question was asked people who did not believe in confidentiality of response stated they had nothing to hide,

so the lack of confidentiality did not deter them from answering. On the attitude survey respondents were asked to complete a self-administered form at the end of the interview. Their answers about things that made them more willing (or less willing) to cooperate indicated that the interviewer's appearance or manner had the most effect on obtaining a response, with a feeling of citizenship also being important. (See Table 14) Although the statement of confidentiality with regard to this study was not as important to respondents as other factors, there was still a sizable number (42 percent) who said that it did make at least some difference in their willingness to be interviewed. As a reason for participation "the topic of the survey" was another "also ran" but again a sizable number (49 percent) said it made at least some difference. Apparently few respondents (2 percent) found it objectionable or uninteresting enough to be a disincentive to participation. If finding the topic objectionable can be taken as an indication of privacy concerns, it does not appear that maintaining privacy was an issue for most respondents on the attitude survey. However, respondents may be dealing with both privacy and confidentiality in a personal sense. The importance respondents attribute to the interviewer's appearance and manner suggests that they were trying to judge whether or not they could trust the interviewer. If the interviewer is perceived as a person who can be trusted to respect privacy and treat answers confidentially, then the respondent may resolve his concerns about these issues without the benefit of prepared statements. To the extent that this personal interpretation is correct, privacy and confidentiality are more important concerns than reactions to guarantees of confidentiality reveal. Other insights into respondent motivations to participate were obtained in the reports of respondents to the attitude survey who had been previously contacted by other surveys. About half of all respondents (54%) reported survey contacts of any kind in the last 4 or 5 years, although not all of these reported contacts may have been bona fide surveys. Reasons for responding or not responding to these contacts are scattered and vary by survey topic and data-gathering mode (mail, telephone, or personal interview). The reasons cited most often for not responding are that the topic was objectionable or uninteresting or the respondent did not want to bother or was too busy.

A multivariate analysis might show that different subgroups of the population are motivated by different combinations of factors. Census and other survey organizations may have to rely on all of these to improve survey response rates.

FOOTNOTES

[1]Some evidence based on recurrent surveys with fairly constant procedures and content is available. For example, see [9].

[2]Note that the views and analysis in this paper are the work of the authors, and do not necessarily reflect the views of the Panel.

[3]The estimator and its variance, the latter derived from Cochran [3] in his discussion of ratio-to-size estimates, are given respectively, by

$$x'_{1A} = 2 \sum^{n_1} \frac{n_1 m_{i1} x_{i1A}}{\Sigma m_{i1}} \quad \text{and}$$

$$\text{Var } x'_{1A} = \frac{n_1^2}{M_1^2} \sum^{n_1} m_{i1}^2 \left(x_{i1A} - \frac{\sum^{n_1} x_{i1A} m_{i1}}{\sum^{n_1} m_{i1}} \right)^2$$

where x'_{1A} is the estimated number of refusals for treatment A units from stratum 1 and Var x'_{1A} is its variance, 2 represents the differential weighting required for stratum 1 as opposed to stratum 2, n_1 is the number of sample clusters in stratum 1, m_{i1} is the number of eligible units in the i-th cluster of stratum 1, M_1 is $\sum^{n_1} m_{i1}$, and x_{i1A} is unweighted value (0,1) of the Treatment A unit in the i-th cluster of stratum 1. Estimators for other treatments and for stratum 2 are defined similarly. The estimated covariance between any 2 treatments was also computed in order to find the standard error of the difference. The between PSU component of variance is not taken into account by the estimator; thus the variances estimated are conditional upon the particular set of 20 PSUs used in this experiment.

REFERENCES

[1] American Statistical Association, "Report on the ASA Conference on Surveys of Human Populations," The American Statistician, 28 (February 1974), 30-34.

[2] Bryant, E.C. and Hansen, M.H., "Invasion of Privacy and Surveys: A Growing Dilemma," in H.W. Sinaiko and L.A. Broedling, eds., Perspective on Attitude Assessment: Surveys and Their Alternatives, Washington, D.C.: Manpower Research and Advisory Services, Smithsonian Institution, 1975, 77-86.

[3] Cochran, W.G., Sampling Techniques, 2nd ed., New York, John Wiley and Sons, Inc., 1963, 300-302.

[4] Conover, W.J., Practical Nonparametric Statistics, John Wiley and Sons, Inc., New York, 1971, 391.

[5] Frankel, Lester R., "Statistics and People--The Statistician's Responsibilities," Journal of the American Statistical Association, 71 (March 1976), 9-16.

[6] Kanuk, Leslie and Berenson, Conrad, "Mail Surveys and Response Rates: A Literature Review," Journal of Marketing Research, 12 (November 1975), 440-453.

[7] Kendall, M.G., Rank Correlation Methods, 2nd ed., Charles Griffin, London, 1955.

[8] Lipset, Seymour M., "The Wavering Polls," The Public Interest, No. 43 (Spring 1976), 70-89.

[9] Love, Lawrence T. and Turner, Anthony G., "The Census Bureau's Experience: Respondent Availability and Response Rates," in Proceedings of the Business and Economic Statistics Section, 1975, Washington, D.C.: American Statistical Association, 76-85.

[10] Market Research Society, "Response Rates in Sample Surveys: Report of a Working Party of the Market Research Society's Research and Development Committee," Journal of Market Research Society, 18 (1976), 113-142.

[11] Reinhold, Robert, "Polling Encounters Public Resistance," New York Times, October 25, 1976.

[12] Snedecor, G.W. and Cochran, W.G., Statistical Methods, 6th ed., Ames, Iowa: The Iowa State University Press, 1967, 246-47.

Table 1. DESIGNED EXPERIMENT: RESPONSE RATES BY TREATMENT

			Treatment Type				
		All Cases	A	B	C	D	E
Total Sample		4420	884	884	884	884	884
No-One-Home Nonresponse -	Total	159	20	22	40	31	46
	Rate	3.6%	2.3%	2.5%	4.5%	3.5%	5.2%
Adjusted Sample (Total Less No-One-Home)		4261	864	862	844	853	838
Refused Before Statement Read -	Total	123	29	26	21	28	19
	Rate	2.9%	3.4%	3.0%	2.5%	3.3%	2.3%
Readjusted Sample (Total Less No-One-Home and Refusals Before Statement)		4138	835	836	823	825	819
Refused After Statement Read	Total	95	15	16	19	22	23
	Rate	2.3%	1.8%	1.9%	2.3%	2.7%	2.8%

Table 2. DESIGNED EXPERIMENT: TEST FOR A LINEAR TREND IN THE PROPORTION REFUSED BY TREATMENT

Treatment	X_i	Weighted a_i	Weighted n_i	$p_i=a_i/n_i$
A	3	15	835	.0180
B	2	16	836	.0191
C	1	19	823	.0231
D	0	22	825	.0267
E	-1	23	819	.0282
		95	4138=N	.0230=$\bar{p}$

$$b = \frac{\Sigma a_i X_i - (\Sigma a_i)(\Sigma n_i X_i)/N}{\Sigma n_i X_i^2 - (\Sigma n_i X_i)^2/N} = -0.00278$$

$$s_b \doteq \sqrt{\frac{\bar{p}\bar{q}}{\Sigma n_i X_i^2 - (\Sigma n_i X_i)^2/N}} = .00165$$

$$t = b/s_b = -1.69 \quad P = .10$$

Table 3. DESIGNED EXPERIMENT: SPEARMAN RANK CORRELATION COEFFICIENT

Measure of the Degree of Concordance Between Hypothesized and Observed Ranks of Treatment Refusal Rates

Treatment	Rank (Hypothesized)	Rank (Observed)	difference	d^2
A	1*	1	0	0
B	2	2	0	0
C	3	3	0	0
D	4	4	0	0
E	5	5	0	0
			$\Sigma d=0$	$\Sigma d^2=0$

*Lowest refusal rate

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - 0 = 1 \text{ (complete concordance)}$$

For n=5, 5% level of significance for r_s, according to Kendall [7], is 1.000. Hence we can conclude that the correlation between the hypothesized and observed rankings is significant.

Table 4. DESIGNED EXPERIMENT: KENDALL'S τ

Measure of Degree of Concordance Between Hypothesized and Observed Ranks of Treatment Refusal Rates

Treatment	A	B	C	D	E
Hypothesized Rank	1	2	3	4	5
Observed Rank	1	2	3	4	5

$$\tau = \frac{N_c - N_d}{n(n-1)/2} = \frac{10-0}{5(4)/2} = \frac{10}{10} = 1 \text{ (complete concordance)}$$

where N_c denotes the number of concordant pairs of observations from the total of $\binom{n}{2}$ possible pairs. N_c is obtained by taking each ranked value for the observed rankings and counting how many ranks to the right of it are greater than it, and adding these counts. N_d denotes the number of discordant pairs. From Conover's Table 11 [4] the critical level for the test statistic ($N_c-N_d = 10$) is estimated to be about $\hat{\alpha} \tilde{=} .01$. Hence we can conclude that the correlation between the hypothesized and observed rankings is significant.

Table 5. Designed Experiment

Do you happen to remember the statement I read at the beginning of this interview? (weighted n = 884 for each treatment)

	Percent of total for each treatment			
Treatment	Noninterview	Yes	No	DK or Other
A	7	82	10	1
B	7	70	11	2
C	9	79	11	1
D	9	75	15	1
E	11	74	15	-

Table 6. Designed Experiment

Did you happen to note whether[1] confidentiality was promised by the Census Bureau? (For Yes answers to above)

	Percent of total for each treatment			
Treatment	Noninterview or not asked	Yes	No	DK or Other
A	18	77	5	-
B	19	76	4	1
C	20	76	3	1
D	24	40	35	1
E	26	50	24	-

[1]It is possible that some respondents anticipated the next question and answered in terms of what the promise was, rather than whether or not they had noted a promise. Thus, for example, a "No" response in condition E may have meant "No, I noted that confidentiality was not promised," rather than "No, I did not note whether confidentiality was promised."

Table 7. Designed Experiment

Was it promised? (Asked of Yes to above)

	Percent of total for each treatment			
Treatment	Not Asked	Yes	No	DK or Other
A	23	76	1	-
B	22	75	1	2
C	23	75	1	1
D	59	26	14	1
E	50	22	27	1

Table 8. Designed Experiment

Was there a time limit given? (Asked of Yes to above)

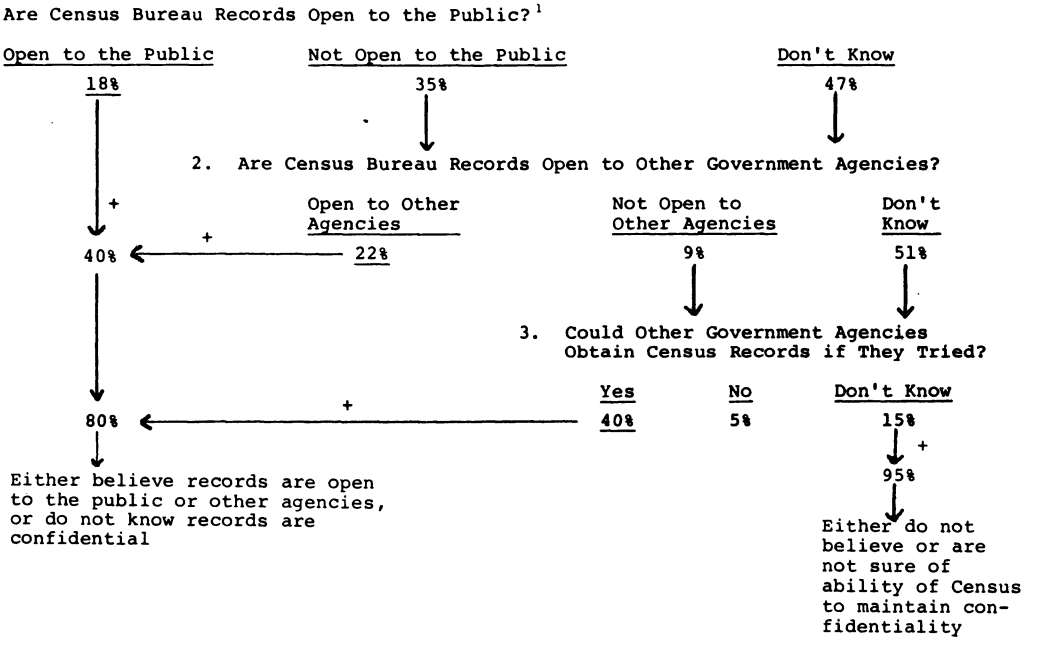
Treatment	Not Asked	Yes	No	DK or Other
A	24	3	71	2
B	23	69	6	2
C	24	67	7	2
D	73	2	23	2
E	77	2	20	1

Table 9. Designed Experiment

What was the limit? (Asked of Yes to above)

Treatment	Not Asked	< 25	25 years	75 years	DK or Other
A	98	-	-	-	2
B	33	1	-	62	4
C	34	3	62	-	1
D	98	1	-	-	1
E	99	-	-	-	1

Table 10. RESPONSES TO QUESTIONS ABOUT CONFIDENTIALITY OF CENSUS BUREAU RECORDS



[1]The specific wording on the questionnaire for these items was, "Individual survey records identified by names and addresses are kept in the files of the United States Bureau of the Census. These records contain information on such things as occupation, income, race and age. Do you happen to know whether these records are public so that anyone who might want to see them can, or are they not open to the public?" For those responding "not open" or "don't know," they were asked, "Do you know whether individually identifiable census records are available to other government agencies or not." For those responding "not open" or "don't know," they were further asked, "Do you feel that other government agencies could obtain individual records from the Bureau of the Census if they really tried?"

Table 11. WHICH TYPE OF ORGANIZATION MOST LIKELY TO GET ACCURATE INFORMATION BY AUSPICES OF COLLECTING AGENT

	Total	National Government	State or Local Government	Universities	Private Companies	Other
Total	100%	40	14	22	10	14
SRC-Michigan	100%	37	11	29	8	16
Census	100%	42	17	16	12	13

Table 12. WHICH TYPE OF ORGANIZATION LEAST LIKELY TO GET ACCURATE INFORMATION BY AUSPICES OF COLLECTING AGENT

	Total	National Government	State or Local Government	Universities	Private Companies	Other
Total	100%	16	6	10	57	11
SRC-Michigan	100%	17	8	4	60	11
Census	100%	15	5	15	54	11

Table 13. REPORT OF WHETHER SOMETHING GOOD OR BAD HAPPENED TO RESPONDENT AS A RESULT OF RESPONDING TO A SURVEY

	Mail	Telephone	Personal
Total	100%	100%	100%
Number of Cases	280	266	201
Yes - Good	10%	4%	10%
Yes - Bad	1	4	1
No	85	88	87
DK/NA	4	4	2

Table 14. **EFFECT OF VARIOUS STIMULI ON WILLINGNESS TO BE INTERVIEWED**

	Survey Sponsorship	Interviewer's Manner	Statement on Confidentiality	Topic of Survey	Curiosity	Sense of Good Citizenship
Total	100%	100%	100%	100%	100%	100%
Much more willing	20	41	23	22	16	31
Somewhat more willing	24	26	19	27	22	33
No difference	45	24	46	41	53	29
Somewhat less willing	1	1	1	1	1	-
Much less willing	1	-	1	1	1	1
Don't know/NA	9	8	10	8	7	6

DISCUSSION

Tore Dalenius, University of Stockholm

The growing difficulties which the authors refer to in the beginning of their paper are not specific to the United States; they have appeared in most other democracies. There is a widespread consensus among survey statisticians that these difficulties are caused to a large extent by the public's concern about invasion of privacy. But it is also clear that our knowledge about the causes is rather scanty.

Consequently, in the last ten years, efforts have been made to get a better understanding of these causes; statistical studies of various kinds have played an instrumental role in these endeavors. The paper just presented is an example in kind.

The paper focusses on areas of prime concern to the Census Bureau and especially its plans for the 1980 censuses of population and housing. It is, however, of a broad scope and should prove useful to most survey statisticians; the data collected represent a most valuable source for action-oriented research aiming at improving the quality of surveys by making their execution more faithful to their design.

In my discussion, I will concentrate on the two studies referred to as "the Behavioral Experiment" and "the Attitude Survey". The designs of these studies both reflect the high competence of those in charge of them, as does the manner in which these designs were implemented. The points of criticism that I will present should not detract from the high appreciation which we should have of these studies.

The Behavioral Experiment

1. The objectives called for testing the effects (if any) on the response rates as well as quality of varying promises of confidentiality. The design properly was one of a comparative experiment with 5 treatments A, B, ..., E in terms of such promises.

It is worth noting that the treatments were verbal stimuli administered by the interviewers as part of the interviews. I suggest that in the final report the authors should discuss the possible effect of these treatments on respondents who prior to the interviews had a conception of the confidentiality of Census Bureau records different from that expressed to them by their interviewers.

2. The analysis of the data is far from final; what is available in the paper represents, I understand, only a minor part of what will appear in the final report.

It is noticeable that the differences in response rates between the five treatment groups are "small". But - as pointed out by our chairman in his opening remarks - even small differences are of great practical significance in the context of the problems likely to be present in the 1980 censuses. Consequently, a seemingly "small" bias may prove serious. As an indication of a possible bias, I refer to the percent "no-one-home": for the group given treatment E it is 5.2%, which is indeed higher than the corresponding percentages for the other four groups.

In the analysis, the authors use two non-parametric procedures. This finds my approval. In addition, they carry out a regression analysis to study the trend in the variable Y = "proportion refused", when the variable X = "treatment" varies in unit steps from X = 3 (the score given for treatment A) to X = -1 (the score given for treatment E). As the authors themselves admit, this scoring is arbitrary. May I suggest that they discard this type of analysis in the final report!

3. The experiment raises an ethical issue which deserves our critical attention. Treatments B, C and E are in fact "misleading"; they misrepresent the policy of the Census Bureau. After the interview, each respondent was informed about the true state of affairs with respect to the promise of confidentiality. We should ask ourselves - as those in charge of the experiment did - if the procedure just described ("temporary deception") is ethically acceptable. I will not pass any judgment of my own here; I want to add, however, that irrespective of which answer we may give to the question, the procedure was a risky one from the viewpoint of the potential harm it might have caused the Census Bureau.

The Attitude Survey

4. The objectives called for measuring attitudes and knowledge about surveys, survey organization, government, confidentiality issues, etc. The design was technically one of a comparative experiment with 2 treatments in terms of the auspices: a government organization and a university organization.

5. Again it is true that the analysis of the data is not final. In my discussion I will focus on three interesting results.

First, the government organization (= the Census Bureau) had a considerably smaller non-response rate than the university organization (= the Survey Research Center at the University of Michigan), mainly due to a smaller refusal rate. This is indeed gratifying to the Census Bureau. A word of warning may nonetheless be in place: according to Brooks and Bailar (1977), the refusal rate in one of the Bureau's key surveys (the CPS) tends to be increasing.

Second - and most surprising to me - the survey indicates that the public is very ignorant about or has a rather low, perhaps dangerously low, opinion about the Census Bureau. Thus 18% of the public thinks that the Bureau's records are open to the public, while 47% do not know if this is the case or not. And many, by far too many, think that the Bureau cannot protect the confidentiality of its records.

Third, the results obtained by the Census

Bureau are in some instances strikingly different from those obtained by the Survey Research Center. Results like these should be kept in mind when we discuss the accuracy of surveys, and especially when we do so on the basis of estimates of the sampling error only.

Some Possible Benefits to the Census Bureau of These Studies

The question whether these studies meet the objectives of the sponsoring agency (= the Census Bureau) should, of course, be answered by that agency itself. This does not preclude, I hope, my discussion of the matter here.

The two findings mentioned before:

i. the positive effect of promises of confidentiality; and
ii. the Census Bureau's poor public image

suggest in my interpretation that the Census Bureau should launch a nationwide "educational" campaign aimed at removing erroneous conceptions and related fears in the public and at enhancing the public's trust in the intentions and capability of the Census Bureau to protect the data collected in surveys and censuses. Just as the Census Bureau has long exercised a leadership in survey and census methodology, it now has the opportunity to exercise a leadership in developing better, much better relations between survey organizations and the public. Action must start now - 1980 is but two years ahead!

Reference

Brooks, C.A. and Bailar, B.A. (1977): An Error Profile: Employment as Measured by the Current Population Survey. Paper presented at the 137th Annual Meeting of the American Statistical Association, Chicago, Ill., August 15-18, 1977.

AN EXPERIMENTAL COMPARISON OF NATIONAL TELEPHONE AND PERSONAL INTERVIEW SURVEYS

Robert M. Groves, University of Michigan

This paper reports a comparison of concurrently administered telephone and personal interview surveys which attempted to collect the same information from national samples of adults. For the 1976 Spring Omnibus Survey, a personal interview sample was contacted by a staff of interviewers dispersed throughout the primary areas of the Survey Research Center's national areal probability sample; concurrently, a telephone sample was called by a group of telephone interviewers centralized in the Ann Arbor offices of SRC. The telephone interview sample was divided into two parts, both containing randomly-generated telephone numbers; one a stratified random sample of telephone households, the other a sample of telephone subscribers in the primary areas of the SRC national sample. The latter design was a feasibility test for mixed-mode surveys that would follow telephone interviews with a personal visit. The two questionnaires included identical attitudinal items on consumer finances, political affairs, relations between the races, and life satisfaction, as well as several factual items.

The discussion summarizes a large group of analyses on the data and compares the two designs on their coverage of the U.S. household population, achieved response rates, ease of obtaining interviews, demographic characteristics of respondents, differences in responses on identical questions, estimates of sampling and interviewer variance, and costs of the data collection.

1. Coverage of the U.S. Household Population by the Two Modes of Surveys

When areal probability methods are applied, errors of field listing do occur, and some members of the population are not covered by the resulting frame. For the SRC national sample of dwellings, undercoverage is estimated to include about five percent of all dwellings in coterminous United States (see Kish and Hess (1958) for a more detailed discussion of noncoverage in areal probability samples).

With random generation of telephone numbers, households in a telephone sample are identified only through their telephone numbers. If a household does not subscribe to telephone service, none of its members can be selected into the sample. The undercoverage in telephone surveys thus is concentrated in a very well-defined subpopulation. In preparation for this project we inserted a question about telephone subscription into the 1975 Fall Omnibus Survey, a national personal interview survey. We repeated that question in this project's personal interview survey and have combined the data to estimate the proportion of households that are not telephone subscribers. Table 1 shows that 7.2 percent of the households are not telephone subscribers. We emphasize that this is 7.2 percent of the respondent households; both surveys are subject to about 25 percent nonresponse. If the nonrespondent households were disproportionately nontelephone households, then our estimate of undercoverage would be low. We were sensitive to this problem and asked interviewers to record on a nonresponse form whether they were able to determine whether or not the household had a telephone. Many times the interviewers found that this was an impossible task, sometimes they made guesses about the existence of a telephone, and other times they determined this with certainty, either by observation or by asking a household member. If the nonresponse data obtained are added to those results, the percentage of households with telephone is largely unchanged.

Despite these efforts at measurement, we prefer a different data source for an estimate

Table 1

Household Telephone Ownership
by Various Household Characteristics
Combined 1975 Fall and 1976 Spring Omnibus Data

	HOUSEHOLDS WITHOUT TELEPHONE	HOUSEHOLDS WITH TELEPHONE	N
1. TOTAL SAMPLE	7.2%	92.8%	3061*
2. REGION			
Northeast	5%	95%	641
North Central	5	95	860
South	13	87	979
West	4	96	581
3. TYPE OF PRIMARY AREA			
Self-Representing Central Cities	9%	91%	234
Suburbs of Self-Representing	1	99	477
Non-Self-Representing SMSA's	6	94	1316
Non-Self-Representing Non-SMSA's	11	89	1034
4. NUMBER OF ADULTS IN HOUSEHOLD			
1 Adult in Household	12%	88%	767
2 Adults in Household	6	94	1859
3 Adults in Household	4	96	312
4 or more Adults in Household	2	98	123
5. NUMBER OF CHILDREN IN HOUSEHOLD			
0 < 18 years in Household	7%	93%	1687
1 < 18 years in Household	7	93	495
2 < 18 years in Household	7	93	465
3 < 18 years in Household	10	90	234
4 or more < 18 years in Household	10	90	176
Missing Data			4
6. RACE			
White	6%	94%	2661
Black	18	82	303
Other	12	88	86
Missing Data			11
7. 1974 FAMILY INCOME			
< $4000	20%	80%	391
$4000 - 7499	13	87	445
$7500 - 9999	10	90	283
$10000 - 14999	4	96	571
$15000 - 19999	3	97	437
$20000 - 24999	2	98	261
$25000 and over	1	99	297
Missing Data			376
1975 FALL OMNIBUS DATA ONLY			
8. HOUSING OWNERSHIP			
Home Owners	4%	96%	948
Renters	17	83	419
Neither Own nor Rent	3	97	37
Missing Data			112
9. HOUSE VALUE FOR OWNERS			
< 15000	16%	84%	161
15000 - 24999	3	97	185
25000 - 34999	2	98	167
35000 or more	0	100	318
Missing Data; Renters			685
10. MONTHLY RENT FOR RENTERS			
$50 or less	28%	72%	58
$51 - 100	26	74	119
$101 - 150	16	84	122
$151 or more	4	96	126
Missing Data; Owners			1091
1976 SPRING OMNIBUS DATA ONLY			
11. TYPE OF STRUCTURE			
Single Family House	5%	95%	1106
Other One Unit Structure	0	100	15
2-4 Total Housing Units in Structure	14	86	157
5-9 Total Housing Units in Structure	16	84	67
10 or more Total HU's in Structure	6	94	101
Trailer in Mobile Home Park	9	91	33
Trailer in Other Location	20	80	55
Missing Data			4

* 6 households of the two sample total of 3067 had missing data on the telephone ownership questions.

of the undercoverage of households by telephones. The Law Enforcement Assistance Administration National Crime Panel study interviews large samples of households each month. Response rates in the study greatly exceed those that SRC studies are able to reach. The January, 1976 panel of the survey contained about 10,000 households, 90.4 percent of which had telephones within the housing unit (Klecka, 1976). We think that this estimate of telephone coverage more accurately describes the problem faced by telephone surveys.

The ten percent noncoverage of households is double that experienced in areal probability samples, and the biasing effects of this noncoverage may be even greater because the households without telephones have very different characteristics from those with telephones. The various subtables of Table 1 show that nonphone households are disproportionately low-income, rural, rented units, likely to contain only one adult, and more likely occupied by blacks than other racial groups. The most important correlate of telephone ownership appears to be family income; telephone samples will fail to include lower income groups in their proper proportions.

The use of telephone surveys alone to infer to the entire household population is inappropriate to the extent that this undercoverage biases sample statistics. For some studies (e.g., surveys of welfare recipients) low income groups are an important portion of the population of interest, and the bias in sample statistics of a telephone survey would be large. For other purposes, when a large proportion of low income groups are not part of the study population, the bias inherent in studying only telephone households would be smaller.

2. Response Rate Analysis

Previous comparisons of personal and telephone surveys have often shown higher response rates for the telephone survey than for the personal interview portion (see Ibsen and Ballweg, 1974). Our experience has generally been the opposite. In this study the response rate for the telephone survey lies between 59% and 70% and for the personal interview survey at 74.3%. The response rate for the telephone survey is presented as a range (see Table 2) because a large group of numbers continually rang without answer when dialed. There was no way to determine whether or not these were working household numbers. The lower telephone response rate counts these as noninterview cases; the higher rate excludes them as noneligible numbers. Later work has shown that the vast majority of these numbers are nonworking, and it is likely that the true telephone response rate is close to 70 percent.[1]

Although the overall personal interview response rate exceeds that of the telephone survey, there are subsets of the population which seem to be accessed more successfully on the telephone. Traditionally, the lowest personal interview response rates are found in the largest metropolitan areas; in the twelve largest SMSA's (all primary areas of the SRC sample) the telephone interview response rate exceeds that of the personal interview (65.5 percent to 61.6 percent). Metropolitan telephone surveys may be

Table 2

Response/Nonresponse Components for Total Telephone Sample

Disposition	n	Percentages Including Ring, No Answers	Percentages Excluding Ring, No Answers
Complete Interviews	1,618	58.6%	70.4%
Partial Interviews	116	4.2	5.0
Refusal by R	203	7.4	8.8
Refusal by Other HU Member	133	4.8	5.8
Non-interview (Other)	208	7.5	9.0
R absent	21	0.8	0.9
Ring, No Answer	460	16.7	99.9%
	2,759	100.0%	

relatively more attractive than personal surveys in those areas. In addition, although the overall telephone response rate is near 70 percent, the rate for the state of Michigan, the area closest to the telephone interviewing staff, is near 80 percent. This result suggests that local telephone surveys, where the sample may have some familiarity with the research organization, may more successfully obtain interviews.

4. Characteristics of Respondents

An examination of the demographic characteristics of respondents may provide some insight into the sources of nonresponse differences between modes.[2] Differences in the distribution of respondents' race, sex, and occupation are negligible or have no clear pattern. Respondents' age and education and total family income, however, reveal consistent discrepancies between the two surveys. A larger proportion of telephone respondents are less than forty-five years of age (Table 3, 60.2 percent) than personal interview respondents (52.3 percent). Larger proportions of telephone respondents (Table 4, 76.3 percent) than personal interview respondents (70.5 percent) failed to obtain a high school diploma. Similarly a larger percentage reported total family incomes of greater than $15,000. In short there is some evidence that younger persons

Table 3

Age of Respondent by Sample Type Using Weighted Data [a]

Respondent Category	Phone	Personal (Households with phone)	Personal (Households with no phone)	Total Personal
18-24 years	16.2%	15.1%	31.5%	16.0%
25-29 years	14.0	12.0	15.2	12.2
30-34 years	10.3	9.5	11.2	9.6
35-39 years	10.3	8.2	6.7	8.1
40-44 years	9.4	7.5	7.9	7.5
45-49 years	7.9	9.5	5.6	9.2
50-54 years	7.8	7.8	5.6	7.7
55-59 years	7.2	7.7	5.6	7.5
60-64 years	6.1	6.8	6.7	6.8
65-69 years	4.7	6.0	2.2	5.7
70-74 years	3.2	4.9	0.6	4.7
75-79 years	1.4	2.6	0.6	2.5
80-84 years	0.9	1.4	0	1.3
85-89 years	0.5	0.8	0.6	0.8
90-94 years	0	0.2	0	0.2
95 or more	0	0.1	0	0.1
TOTAL	99.9%	100.1%	100.0%	99.9%
% Unweighted N	1575	1421	106	1527
MISSING DATA				
Terminated	103			
Other	56	18	3	21

[a]Data weighted by reciprocal of selection probability

Table 4

Education Summary of Respondent by Sample Type Using Weighted Data[a]

Respondent Category	Phone	Personal (Households with phone)	Personal (Households with no phone)	Total Personal
8 grades or less	8.2%	12.7%	34.1%	14.0%
8 grades or less, plus non-academic training	1.3	1.7	1.1	1.6
9 - 11 grades, no diploma	10.7	11.5	21.8	12.1
9 - 11 grades, no diploma, plus non-academic training	3.5	3.6	5.0	3.7
High School diploma	21.6	21.6	17.9	21.4
High School diploma, plus non-academic training	14.1	12.9	10.6	12.8
Some college - 1/2 year - 3 years	22.3	19.5	6.1	18.7
Junior or Community college degrees	1.6	1.9	2.2	1.9
BA level degrees	11.9	10.7	0.6	10.1
Advanced degree including LLB	4.9	3.9	0.6	3.7
Don't Know	0	0	0	0
TOTAL				
%	100.1%	100.0%	100.0%	100.0%
Unweighted N	1607	1431	107	1538
MISSING DATA				
Terminated	103			
Other	24	8	2	10

[a]Data weighted by reciprocal of selection probability

5. Response Differences Between Modes

A comparison of response distributions from the two modes in this project can suggest topic areas or question types that may be better measured in one mode than the other. We cannot estimate the pure effect of administration mode because two different interviewing staffs conducted the surveys, because each survey is subject to its own nonresponse problems, and each survey covers different portions of the U.S. household population. The latter complication can be alleviated by comparing the telephone respondents with those personal interview respondents in telephone households. Even with this control, however, we can only contrast two bundles of methodologies, each with its own collection of errors and effects of administrative organization.

Over two hundred different measures common to both modes were obtained; only a few statistically significant differences between modes were obtained. Some differences that are visible suggest weaknesses in the telephone survey data. Missing data due to failure of the respondent to answer or of the interviewer to ask the question were found to be somewhat higher on the telephone than in face-to-face interaction. On later SRC telephone surveys asking the same questions, we found that the missing data rate on the telephone survey declined over time to very near that of the personal interview survey. The result supports the hypothesis that a telephone interviewing staff can improve with experience.

Another weakness in the telephone survey data appears on open-ended items where fewer respondents offer several different thoughts in response (see Groves, 1976). One question was inserted in both questionnaires specifically to investigate this problem. A list of important problems facing the country was requested, and the probing to be used by interviewers was written into the instrument. About eleven percent fewer telephone respondents than personal interview respondents supplied three or more problems. In multivariate analysis of this measure, younger, more affluent respondents and those judged more interested in the interview were found to exhibit the largest differences between mode. When it was noted that the telephone interviews generally were faster paced, the conjecture was made that these groups, who often supply full and detailed answers, might more quickly adjust their behavior to the faster pace.

Another indicator of potential problems in the telephone survey data arose from attitudinal measures gauging the respondent's reaction to the interview. Fewer telephone respondents (39.4 percent) preferred that mode of answering questions (relative to face-to-face or self-administered questionnaires) than did personal respondents prefer the face-to-face mode (78.4 percent). Proportionately more telephone respondents noted that they felt "uneasy" about discussing some topics, especially their financial status and political attitudes. The telephone interviewers observed more suspicion and questions about the legitimacy of the study than did personal interviewers.

Other differences that exist do not suggest weaknesses in one of the modes but rather the effects of varying constraints in the two modes. Questions utilizing response cards in the face-to-face interviews were adapted to the telephone in a variety of ways. We found that the differences between modes on these questions seem to be sensitive to how many points on the scales are labelled, whether the scale is numerically-based (e.g., income, years of education). Method effects also depend on whether the telephone interviewer presents the entire scale or first its major categories (e.g., agree, disagree) followed by more specific categories (e.g., strongly, weakly disagree).

We found little evidence of different responses to items with socially desirable answers (see Hochstim, 1967; Colombotos, 1965). Although there is some evidence of greater respondent optimism on the telephone for consumer sentiment items and life satisfaction items, later surveys suggest that this was not a reliable result. Consistent with past results (Rogers, 1976), negligible differences between modes were found on reports of voting behavior.

Although we found few differences between mode on the total sample, many analyses on such data use statistics calculated on subclasses. Using age, education, income, and race groups, we searched for subsets of the population that might reveal differential effects of mode. This was largely unsuccessful; the differences were usually within sampling error and somewhat unstable across measures.

6. Calculation of Sampling Errors

In all three of the sample designs used in this project, sampling variance arises from two different sources, differences among persons that happen to be selected on different draws of the sample and differences of sample size achieved in different draws. In addition, random-digit dialed samples experience sample size variation because they search for a subset of all ten-digit telephone numbers. There is no control on what proportion of sample telephone numbers are working

Table 5

Sampling Error Calculations for Stratified Phone, Clustered Phone, and Total Personal Interview Samples

Variable Description	Mean Value or Proportion of Adults			N			Square Root of Design Effect				Coefficient of Variation		
	Stratified	Clustered	Personal	Stratified	Clustered	Personal	Stratified	Clustered	Personal	Personal Reduced	Stratified	Clustered	Personal
Reporting they live in a rural area	.19	.21	.34	790	829	1548	1.10	1.20	.93	.96	.0376	.0354	.0259
Reporting they live in or near a city of 50,000 or more	.39	.38	.66	720	759	1548	1.07	1.13	.93	.96	.0394	.0351	.0259
Reporting that they itemized deductions on 1975 tax return	.53	.53	.47	750	789	1486	1.07	1.11	.95	.97	.0387	.0358	.0286
Feeling Satisfied to Completely Satisfied about life as a whole	.84	.83	.83	402	401	723	1.02	1.27	1.00	1.00	.0533	.0660	.0341
Reporting total family income less than $7,500	.19	.20	.26	662	703	1348	1.00	.80	1.05	1.03	.0414	.0378	.0334
Feeling Mostly Satisfied, Pleased, or Delighted about life as a whole	.80	.79	.86	393	435	811	1.05	1.09	1.09	1.04	.0538	.0505	.0325
Feeling better off financially now than 1 year ago	.38	.38	.36	837	865	1531	1.07	1.23	1.09	1.05	.0366	.0456	.0252
Reporting that they planned to vote in 1976 Presidential Election	.85	.86	.78	780	796	1548	1.04	1.28	1.11	1.06	.0379	.0390	.0259
Reporting that they were not presently working	.37	.35	.42	799	838	1547	1.07	.97	1.19	1.10	.0374	.0369	.0258
Feeling saving money more important now than usual	.64	.59	.68	800	837	1494	1.05	1.46	1.20	1.11	.0374	.0451	.0264
Reporting that they voted in 1972 Presidential Election	.65	.70	.62	787	814	1507	1.12	1.32	1.18	1.10	.0377	.0389	.0269
Feeling "Very Happy" these days	.34	.30	.30	794	821	1521	1.07	1.41	1.21	1.12	.0376	.0412	.0257
Not obtaining at least a high school diploma	.22	.25	.31	779	827	1538	1.07	1.37	1.21	1.12	.0379	.0345	.0262
Mean feeling thermometer rating for Gerald Ford	52.90	52.96	54.29	734	769	1485	1.02	1.00	1.22	1.13	.0391	.0428	.0271
Who are 18-29 years old	.31	.30	.28	769	806	1527	1.12	1.06	1.26	1.15	.0382	.0343	.0253
Thinking of themselves as a Democrat	.49	.53	.53	759	794	1516	1.08	1.25	1.28	1.16	.0386	.0357	.0260
Feeling Whites have right to keep Blacks out of their neighborhood	.06	.07	.10	784	812	1525	1.06	1.41	1.34	1.19	.0378	.0376	.0262
Mean feeling thermometer rating for Jimmy Carter	54.57	55.26	57.53	616	630	1290	1.05	1.18	1.46	1.31	.0430	.0463	.0309
Mean number of telephones in home	1.89	1.92	1.73	800	838	1546	.78	1.00	1.54	1.31	.0374	.0375	.0258
Mean number of problems facing the country	3.99	4.02	4.28	775	826	1535	1.06	1.22	1.61	1.35	.0380	.0397	.0261
Who are nonwhite	.13	.13	.14	782	818	1545	1.06	.99	1.62	1.36	.0378	.0342	.0260
Feeling Cockroaches are not a problem in their home	.73	.76	.75	798	836	1546	1.07	1.37	1.74	1.44	.0374	.0372	.0258
Mean over 22 variables							1.05	1.19	1.24	1.12			

household numbers. In this project about 22 percent of all sample numbers were household subscriptions, but other samples could have by chance experienced a higher or lower proportion of eligible numbers. This source of variation in sample size is present in both telephone samples. Finally, the sample size of the clustered telephone design varies for one additional reason. Some telephone exchanges serve both households within and outside a primary area of the SRC national sample. Telephone numbers selected from these exchanges were screened, and in total we found that about seventy percent of them serve households within the primary area. Unfortunately, there is no control on this proportion and it could vary over different sample draws creating different totals of eligible household numbers generated.

Table 5 presents sampling errors for the statistics calculated on the total sample.[3] All statistics are proportions of the total sample except for those that are labelled as mean values. We present four separate pieces of information for each sample type: the mean value or proportion of adults having such a characteristic, the unweighted number of observations, the square root of the design effect, and the coefficient of variation of cluster size. All means and proportions are calculated using the selection weights arising from variation in number of eligible respondents in the sample household. The design effect, deff, is presented as a measure of the relative precision of the means and proportions. The square root of deff (called deft) is the ratio of the two standard errors. Since many packaged computer analysis programs produce estimates of variances or standard errors based on the assumption of simple random sampling, deff's or deft's can be used as multiplicative adjustments to these values to calculate the appropriate sampling error or to adjust confidence intervals to account for the complexities of the sample design. By comparing the variance of the design to that of a simple random sample of the same size, deff's also adjust for differences in the number of interviews in each sample. For the stratified random telephone sample, a deft greater than 1.0 or increased variance relative to a simple random sample of the working household numbers arises from the lack of control of sample size, and for the clustered telephone sample, from both lack of control on sample size and clustering effects. For the personal interview sample, deft's greater than 1.0 arise from the effects of clustering.[4] We expect the deft's for the stratified random telephone sample to be lower than those of the clustered telephone sample for the same statistic.

The final section of Table 5 presents coefficients of variation for the cluster size in the three different designs. All three samples have coefficients of variation safely below the level threatening the ratio mean variance approximation, (they range from .02 to .05), but the figures do provide evidence for the increased variability in size within the telephone samples (about a 40% increase in the coefficient of variation). This

reflects the variation in proportion of working numbers across the central office codes sampled.

The deft's in Table 5 are arranged by their value within the personal interview sample from lowest to highest.[5] Using the reduced personal sample, the range is .97 to 1.44 with an average over the twenty variables of 1.16. For the clustered telephone sample the order of estimates by the deft values is somewhat different, but the range of values is .80 to 1.46, with a mean deft of 1.19. The stratified telephone sample in general has the lowest design effects, a range from .78 to 1.12, and a mean deft over the twenty proportions of 1.05.

We are reminded by this exercise that although the clustered telephone sample is probably subject to less control over sample size than the personal interview sample in the same primary areas, telephone sampling within primary areas selects elements directly, all over the area, while the personal interview sample further clusters the sample into secondary units. The added clustering within primary areas in the personal interview sample may produce higher design effects than an element sample spread over the entire area. Thus, the effects of lack of control over sample size in the telephone sample may be nearly balanced by the secondary clustering effects in the personal sample.

Comparing the stratified and clustered telephone designs, we observe an average 14 percent increase in the standard error for the clustered sample. That reduced precision added to the forty to fifty percent increase in sample numbers required in the clustered sample makes the clustered design more attractive only for studies planning later personal interviews in the same households or studies of change from estimates obtained in other studies in the SRC primary areas.[6]

7. Interviewer Effects Within the Telephone Survey

One source of nonsampling error can be linked to the interviewers. Past research has demonstrated that individual interviewers may, because of different styles of asking questions, personality differences, or interactions of respondent and interviewer characteristics, produce different responses from the same respondents (e.g., Hanson and Marks, 1958; Dohrenwend et al, 1968). Following the approach of Hansen, Hurwitz, and Madow (1953), we characterize the effect of interviewer differences on the variance of a sample mean or proportion as a design effect:

$$\text{Deff}_{int} = 1 + \rho_{int} \quad (b_{int} - 1)$$

where ρ_{int} is a measure of within-interviewer homogeneity, reflecting the extent to which answers of an interviewer's respondents resemble one another, and where b_{int} is the average number of interviews taken by an interviewer.[7] This design effect measures the change in the variance of sample estimates due to the fact that clusters of respondents were interviewed by the same person instead of by different people. If there are interviewer effects on responses, respondents of the same interviewer will tend to give distinctive answers, ρ will be positive and the $\underline{\text{deff}}_{int}$ will be greater than one.

In order to calculate $\underline{\text{deff}}_{int}$, the interviewers must be selected at random from among those available, and be assigned sample elements at random to eliminate any covariation of interviewer attributes with respondent attributes. Randomized selection of interviewers from among those judged eligible did not occur; indeed the selection process attempted to achieve a uniformly high interviewer quality, and homogeneity, rather than heterogeneity, across interviewers would be the expected result of the personnel decisions. The effect of this departure would presumably decrease interviewer variance and our analysis will probably err on that side. Conversely, in terms of inference to later project experiences, the personnel decisions will probably be repeated, and this project's results are useful guides to later results. The second requirement for estimating $\underline{\text{deff}}_{int}$, the randomization of assignment of sample elements to interviewers, was painstakingly implemented in the project. As part of the sampling process, equal-sized subgroups of the sample were randomly assigned to interviewers so that, in essence, each interviewer was responsible for a small national sample. Since the telephone interviewers worked specific hours within each day, however, they could not make calls on numbers at all hours, and periodically sample numbers were randomly reassigned manually to interviewers that worked different shifts. What results from the process is a randomization within interviewer shift. Because of this, the $\underline{\text{deff}}_{int}$ measured will also contain differences between the types of interviewers that work different shifts and respondents reached during different shifts. We suspect that respondent differences across shifts are largest between those reached on weekday mornings and afternoons on one hand, and those reached on weekday evenings and weekends. An examination of the personnel on each shift shows that about two-thirds of the interviewers work in both of these groups, and we have collapsed over shifts in the analysis that follows.

Values of ρ^*_{int} were calculated for the twenty-four estimates; their values range from -.01 to .07.[8] The highest ρ^*_{int} (.071) corresponds to the number of problems facing the country mentioned by respondents. This number is probably affected by the quality of probing used by the interviewer. We noted earlier that respondent behavior regarding this question seems to differ by mode of interview. Other estimates subject to high interviewer variance are the proportion feeling that it is more important than usual to add their savings (an open ended attitudinal measure, $\rho^*_{int} = .045$) the proportion who report that they are not currently working (a sensitive subject to some respondents, $\rho^*_{int} = .038$), the percentage of respondents who did not reveal their total family income (either directly or by responding to the trichotomous categorization of income, $\rho^*_{int} = .027$). Two estimates arise from the same questions as two of the above but have much lower interviewer effects. The proportion of respondents whose total family income was less than $7,500 has a small positive ρ^*_{int} (.003), and the proportion of respondents who fail to mention any problem facing the country has a small negative ρ^*_{int} (-.001). The discrepancies in interviewer effects between the two estimates related to total family income could

support the hypothesis that reluctance to provide income to the interviewer may result from interviewer inflection or hesitation in asking the question (a variable over interviewers); once committed to giving an income figure, the proportion who reveal a low income (less than $7,500) is rather stable over interviewers. The questions asking for a listing of the most important problems facing the country should have a different pattern; we would expect relatively large interviewer effects both for the mean number of problems mentioned and the proportion of respondents who cannot identify any problems. The former is highly variable over interviewers (ρ^*_{int} = .071), but the rate of "don't know" on the item is fairly stable (ρ^*_{int} = .001). It may be the case that initial delivery style of the question has little effect on the probability of a respondent mentioning at least one important problem. In contrast although the probing was specified in the questionnaire, the number of problems mentioned seems to be much more dependent on interviewer style.

These results inform us about interviewer effects in this telephone survey, but we cannot present a comparable analysis for the personal interview survey. Despite this, a comparison of telephone interviewer effects with those of previous personal interview surveys may give some insight into the relative magnitudes of interviewer variance in the two modes. To do this we utilize three published studies: 1) Hanson and Marks' (1958) analysis of enumerator variance in 21 counties of Ohio and Michigan during the 1950 U.S. Population Census, 2) Kish's (1962) study using two surveys of factory workers, and 3) Freeman and Butler's (1976) study using a survey of urban housewives.

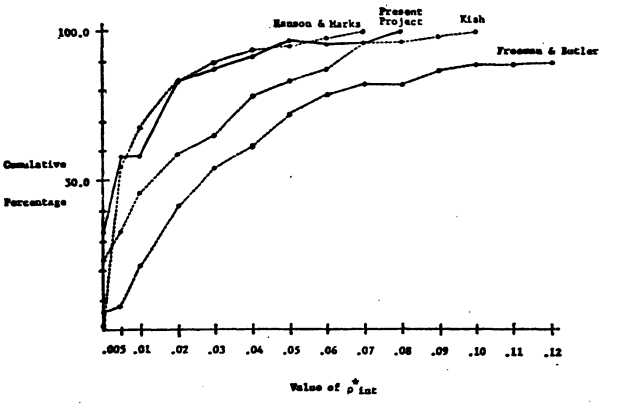
The census study includes purely demographic measures, some sensitive like income, others without any threat to the respondent, like sex; still others measure missing data on schedules returned by the interviewers. The Kish study yielding the largest range of ρ's, asked attitudinal questions about union activities and job satisfaction, in addition to some purely demographic measures. The Freeman and Butler study calculated ρ's on all categories of seventeen different variables, some of them attitudinal variables related to the general topic of mental retardation among children, others are reports of their actions toward their own children, or reports on personal behavior of other kinds.

Interviewers in the Census study were those paid as enumerators in that Census, Kish's studies used professional male interviewers employed by the Survey Research Center, and the Freeman and Butler study used school teachers none of whom had interviewing experience, but who participated in a "three-credit-hour university course in interviewing conducted by the project and field directors" (p. 84).

Figure 1 presents cumulative percentages for values of ρ's and ρ^*_{int}'s for the four different studies. The results of the telephone study are plotted with the solid black line. The highest ρ's are those found by Freeman and Butler's study of housewives. The Census study has the smallest ρ's, although our telephone survey produces the largest proportion of ρ^*_{int}'s less than zero. The Freeman

Figure 1

Cumulative Percentage Distribution for ρ^*_{int} Values Measuring Interviewer Effects for Three Personal Interview Surveys* and this Telephone Survey



* Hanson and Marks (1958) and Kish (1962), Study 1, distribution taken from Table 2, p. 97 of Kish. Freeman and Butler (1976) distribution taken from their Table 1, pp. 86-87.

and Butler study exhibits interviewer effects much higher than any of the other studies, and the use of new, nonprofessional interviewers may be associated with that result. Even ignoring that result, however, it appears that the interviewer variance experienced in the telephone survey is often lower than that in the personal interview surveys included in Figure 1.

Although the inference from Figure 1 is complicated by variation in type of measures, interviewers and populations, it suggests that telephone interviewer effects measured by ρ^*_{int}'s may be somewhat smaller than those in personal interviews. The important lesson of Figure 1, however, requires additional information. As we noted earlier, the effect on the variance of sample estimates corresponding to interviewer differences can be characterized as:

$$deff_{int} = 1 + \rho^*_{int}(b_{int} - 1)$$

where b_{int} is the average number of interviews taken by an interviewer. We have presented ρ^*_{int}'s in order to control differences in the workload of interviewers across the different studies. This is a proper approach when comparing the magnitude of interviewer variation in the two modes, but it ignores possible administrative differences in the modes. In the telephone survey interviewers each completed an average of forty-four interviews; the corresponding number in the personal interview survey is eleven. With a ρ^*_{int} of .04, which is likely in both surveys for an open-ended or sensitive item, the $deff_{int}$ for the telephone survey is 2.72; for the personal, 1.40. Simply because the telephone interviewers each take more interviews, the loss of precision arising from interviewer effects is larger. Indeed, the interviewer differences measured by ρ^*_{int} have to be less than one quarter their size in the personal interview survey for the design effects due to interviewer differences to be the same. The results in Figure 2 suggest that this will not always be the case. This illustrates that interviewer effects within centralized telephone interviewing facilities may be a larger threat to survey precision than in dispersed personal

interviewing situations. The very fact that all telephone interviewers work in the same location, and that there are relatively few of them, however, facilitates the study of methods to reduce interviewer variance in ways not possible in personal interviewer studies.

The data on sampling and interviewer variance should be combined to provide estimates of change in standard errors of the telephone survey as we administered it from one yielding the same estimates from a simple random sample interviewed singly by different interviewers.[9] The columns in Table 5 listing the square roots of design effects and ρ's for sampling and interviewer differences can be used to provide an overall effect. Table 6 presents an ordering of the overall deft's for the twenty-two estimates common to the sampling error and interviewer variance analysis separately for the stratified and the clustered telephone samples. The deft's range from .74 to 1.57 in the stratified sample and .83 to 1.75 in the clustered sample. This implies a 60-75 percent increase in the width of confidence intervals for some sample statistics. For those variables sensitive both to clustering and to interviewer effects (e.g., attitudes about the need for saving money), the total design effect is rather large (deft = 1.75), but in some cases the interviewer effects actually decrease the overall design effect from that due to sampling alone. This overall design effect may be a more proper inflation factor for simple random sample standard errors that are produced by most packaged computer programs.

Table 6

Estimated Overall Design Effect Including Sampling and Interviewer Variances for Twenty-two Measures by Telephone Sample Type

Variable Description	Square Root of Design Effects				Square Root of Overall Design Effect	
	Stratified Sample		Clustered Sample			
	Sampling	Interviewer*	Sampling	Interviewer*	Stratified Sample	Clustered Sample
Reporting total family income less than $7,500	1.00	1.02	.80	1.03	1.02	.83
Mean number of telephones in home	.78	.97	1.00	.97	.74	.96
Who are 18-29 years old	1.12	.92	1.06	.92	1.06	.98
Who are nonwhite	1.06	1.04	.99	1.04	1.09	1.03
Mean thermometer rating for Gerald Ford	1.02	1.10	1.00	1.11	1.11	1.10
Mean thermometer rating for Jimmy Carter	1.05	.91	1.18	.91	.97	1.11
Reporting they live in or near a city of 50,000 or more	1.07	.98	1.13	.98	1.05	1.12
Reporting that they itemized deductions on 1975 tax return	1.07	1.02	1.11	1.02	1.09	1.13
Feeling Mostly Satisfied, Pleased, or Delighted about life as a whole	1.05	1.04	1.09	1.05	1.09	1.14
Feeling better off financially now than one year ago	1.07	.99	1.23	.99	1.06	1.22
Feeling Satisfied to Completely Satisfied about life as a whole	1.02	.98	1.27	.98	1.00	1.25
Reporting that they planned to vote in 1976 Presidential Election	1.04	1.04	1.28	1.05	1.08	1.31
Reporting that they were not presently working	1.07	1.32	.97	1.33	1.37	1.32
Reporting they live in a rural area	1.10	1.13	1.20	1.13	1.22	1.32
Thinking of themselves as a Democrat	1.08	1.09	1.25	1.10	1.16	1.33
Not obtaining at least a high school diploma	1.07	.97	1.37	.96	1.04	1.34
Feeling cockroaches are not a problem in their home	1.07	1.01	1.37	1.01	1.08	1.37
Feeling "Very Happy" these days	1.07	1.01	1.41	1.01	1.08	1.41
Reporting that they voted in 1972 Presidential Election	1.12	1.13	1.32	1.13	1.23	1.42
Feeling Whites have a right to keep Blacks out of their neighborhood	1.06	1.17	1.41	1.18	1.22	1.54
Mean number of problems facing the country	1.06	1.53	1.22	1.56	1.57	1.71
Feeling saving money more important now than usual	1.05	1.37	1.46	1.39	1.41	1.75

* Those deft's were estimated by using the ρ^*_{int} values presented in Table 5.4 and the number of valid responses in Table 5.2.

8. Sampling and Data Collection Costs for the Surveys

The small literature that does exist regarding telephone surveys frequently contains references to costs associated with the method. Coombs and Freedman (1964) estimate that using the telephone wherever possible in a reinterview of respondents "resulted in savings of approximately 60 percent." The field cost per five-minute telephone interview of the national sample of Kegeles et al (1969) was about six dollars which they labeled "only a fraction of what a personal interview would cost." Hochstim (1967) incurred telephone interviewing costs which were fifty to seventy percent of those for the same interview completed in person. Tuchfarber and Klecka (1976) estimate personal interview costs at five times the costs for a comparable RDD survey of Cincinnati households.

Before we describe our methods of cost analysis, we should outline several dangers of inference from the costs of any one project. Each survey has unique characteristics which affect its total costs: the nature of the population studied, the size of the sample, the length and complexity of the questionnaire, and the number of interviewers employed. This project itself has some characteristics which may or may not be duplicated in future studies of either mode. This was the first telephone survey with randomly generated sample numbers ever conducted by the Survey Research Center; new methods, however pretested, inevitably bring with them difficulties of administration. Since this project we have completed other such telephone surveys and are enjoying greater efficiency in some areas than we did earlier. Also, because of the methodological nature of this telephone survey, the research staff had a larger involvement in the interviewing process than in later telephone surveys, and its participation no doubt reduced the activities of the field office personnel. Other qualities of the two different surveys, while each typical of the particular mode, may complicate the comparison of costs between modes. For example, the average personal interview lasted about fifty minutes, the telephone, only thirty minutes.[10] All these complications limit the utility of our data to other researchers for judging costs of either survey mode. We have chosen not to adjust costs in the two surveys in an attempt to reduce differences; rather we will present costs actually incurred by the two modes.

Table 7 summarizes the direct costs for sampling and field activities on the two studies. The table is broken into ten categories, representing major divisions of work. Costs for all items, person hours for salary items, and unit counts for non-salary items are listed for the components of each category.

Table 7

Direct Costs for Components of Sampling and Data Collection Activities on the Telephone and Personal Interview Surveys

	Telephone Survey		Personal Interview Survey	
	Hours or Other Units	Costs	Hours or Other Units	Costs
I. Sampling				
Administrative Salaries	86.0*	$ 505.27*	362.0	$3,305.03
Clerical/Typing Salaries	0	0	186.0	676.12
Chunking and Listing		0		4,366.00
Data Processing		450.00		0
Category Total	86.0 hours	$ 955.27	548.0 hours	$8,547.15
Percentage of Total	1.6%	2.5%	4.1%	10.1%
II. Pretest				
Ann Arbor Field Office Salaries	34.0	$ 224.59	32.0*	$ 200.08*
Clerical/Typing Salaries	4.0	14.73	17.0	88.65
Supervisors Salary	50.5	188.67	25.8	153.30
Interviewers Salary	76.4	226.06	125.7	444.13
Travel	0	0	(666mi)	93.32
Duplicating	(1,688p)	69.40	(7,370p)	119.62
Postage		0		14.00
Category Total	164.9 hours	$ 723.45	200.5 hours	$1,113.10
Percentage of Total	3.0%	1.9%	1.5%	1.3%
III. Training and Prestudy Work				
Interviewing Supervisors Salaries	37.1	$ 155.57	667.0*	$3,929.14*
Interviewers Salaries	314.6	936.58	1,621.3*	$5,285.51*
New Interviewer Training		660.00		0
Duplicating	(324p)	26.93	(6,725p)	85.10
Supplies		16.78		223.86
Coding Staff Salaries	11.0	94.46	0	0
Coding Evaluation of Questionnaires	40.0	178.00	0	0
Category Total	402.7 hours	$2,066.34	2,288.3 hours	$9,523.61
Percentage of Total	7.4%	5.4%	16.9%	11.2%
IV. Materials				
Questionnaire	(100,800p)	$ 802.25	(224,000p)	$1,466.51
Other Data Collection Instruments		278.70	(24,840p)	704.65
Data Collection Related Materials and Reporting Forms		274.68	(30,900p)	1,211.24
General Supplies		19.33		277.75
Category Total		$1,374.96		$3,660.15
Percentage of Total		3.6%		4.3%
V. Ann Arbor Field Office				
Administrative Salaries	156.0	$1,222.48	324.0	$2,508.29
Clerical/Typing Salaries	55.0	172.26	392.4	1,651.13
Category Total	211.0 hours	$1,394.74	716.4 hours	$4,159.42
Percentage of Total	3.9%	3.7%	5.3%	4.9%
VI. Field Salaries				
Supervisor Salaries	648.0	$2,303.64	988.5*	$4,956.88*
Interviewer Salaries	3,442.0	10,181.05	8,389.8*	27,321.04*
Foreign Interviewers Salaries	(6int)	60.0	0	0
Category Total	4,090.0 hours	$12,544.69	9,378.3 hours	$32,277.92
Percentage of Total	75.5%	33.1%	69.4%	38.0%
VII. Field Staff Travel				
Supervisor Travel		0		$5,620.35
Interviewer Travel				
Personal Auto Mileage	0	0	(74,405mi)	10,416.72
Other		0		778.04
Category Total		0		$16,815.11
Percentage of Total		0		19.8%
VIII. Communications				
Postage		0		$3,491.03
Telephone				
For Data Collection		$15,793.60		0
For Other Communications		0		1,756.45
Supplies For Mailing		0		732.83
Category Total		$15,793.60		$5,980.31
Percentage of Total		41.6%		7.0%
IX. Control Function				
Administrative Salaries	247.5	$ 830.55	8.0	$ 91.30
Clerical/Typing Salaries	0	0	188.5	766.93
Printing and Duplicating	0	0	(1,500p)	24.99
Data Processing		372.00*		0
Category Total	247.5 hours	$1,202.55	196.5 hours	$ 883.22
Percentage of Total	4.6%	3.2%	1.5%	1.0%
XI. Post Interviewing Activities				
A. Interviewer Evaluation/ Debriefing				
Supervisor Salaries	12.0	$ 44.52	0	$ 0
Interviewer Salaries	68.0	199.70	30.2*	98.62*
Ann Arbor Administrative Salaries	0	0	16.0	94.87
Ann Arbor Clerical/Typing Salaries	0	0	12.0	37.80
Duplicating	(50p)	2.50	(1,200p)	15.55
Postage		0		34.45
Category Total	80.0 hours	$ 245.22	58.2 hours	$ 281.29
Percentage of Total	1.5%	0.6%	0.4%	0.3%
B. Verification				
Ann Arbor Administrative Salaries	$ 100.1*	$ 384.78*	90.5	$ 596.44
Ann Arbor Clerical/Typing Salaries	0	0	9.0	35.13
Duplicating	0	9.30*	(1,150p)	33.90
Supply		24.79*		23.88
Postage		88.66*		104.78
Telephone		0		82.25*
Category Total	100.1 hours	$ 507.53	99.5 hours	$ 876.38
Percentage of Total	1.8%	1.3%	0.7%	1.0%
C. Report to Respondents				
Ann Arbor Administrative Salaries	3.0	$ 22.63	4.0	$ 28.70
Keypunching	34.2	256.23	32.8	246.00
Data Processing		162.32*		112.94
Printing	0	555.00*	(22,400p)	251.00
Postage		133.76*		107.62
Category Total	37.2 hours	$1,129.94	36.8 hours	$ 746.26
Percentage of Total	0.7%	3.0%	0.3%	0.9%
OVERALL TOTAL	5,419.4 hours	$37,939.29	13,522.5 hours	$84,863.92
PER INTERVIEW TOTAL	3.3 hours	$ 23.45	8.7 hours	$ 54.82

*Costs based on estimates of those personnel involved in the work, usually necessitated by different categories of work being performed by the same personnel.

Total direct sampling and field costs for the personal interview survey are $84,864. For the telephone survey, the costs total $37,939, only about 45 percent of those on the personal study. Person-hours total 13,523 on the personal mode, 5,419 on the telephone mode. For these two studies, therefore, the telephone mode is substantially less expensive, both in terms of direct costs and personnel time required. These results resemble those reported by Hochstim (1976) and Coombs and Freedman (1969).

For the two samples the per completed interview cost for sampling and field work is $55 using personal interviews, $23 using telephone interviews. This involves an average of 8.7 person hours per personal interview, and 3.3 person hours per telephone interview. Sample sizes were 1,548 for the personal interview study, 1618 for the telephone interview study.

While we assume that costs in survey areas other than sampling and field should be unaffected by differences in interviewing method, it would perhaps be helpful to consider our figures in the context of total survey costs. Analysis costs probably have the highest variation of all components, but we roughly estimate that sampling and field costs comprise about 50 to 60

percent of personal interview survey direct costs incurred before analysis. Expecting these other activities to cost the same for a telephone survey, we would estimate that 31 to 40 percent of total telephone survey costs up to analysis are attached to sampling and field work. Using these figures, we would expect that the total telephone survey costs would be 56 to 87 percent of total personal interview costs before analysis.

Table 7 identifies areas where large portions of sampling and field costs were incurred in each of the two modes and where large cost differences exist between the two modes. There are five areas that exhibit the largest differences. Sampling, prestudy, and training costs were markedly different in the two modes. Travel costs accounted for nearly 20 percent of total personal interview costs but were nonexistent on the telephone survey. Total communications costs (mainly WATS lines charges), on the other hand, formed over a third of all telephone survey charges and were three times as large as those for the personal interview survey. In both modes, interviewer and supervisor salaries accounted for about a third of all sampling and field costs.

There are two design differences in our studies which complicate cost comparisons. First, the fact that the sample sizes on the two studies are not identical makes use of a per interview cost somewhat difficult. We might wish to estimate costs for a different survey by multiplying the sample size by per interview cost, assuming constant marginal cost of a single interview across different sample sizes. It is more plausible that the cost of taking one interview decreases as the number of interviews increases. Therefore, having a larger telephone sample (N = 1,618) probably yields slightly lower per interview costs than would exist if the telephone sample size were 1,548. However, since the difference between the two sample sizes is small (70 cases) relative to total sample sizes (1,548 personal, 1,618 telephone) the effects of increased size are probably small.

A more serious design difference is the discrepancy in interview lengths on the two studies. To adjust for this difference, we counted the number of variables obtained in each mode. We enumerated non-missing data records on all variables that were the direct result of responses recorded by the interviewer. An approximate count for the personal interview is 289,400 and for the telephone, 260,500.[12] Using these estimates the per unit data costs are about $.29 for the personal and $.15 for the telephone survey (about 50 perent of the personal).

Another approach to calculating per unit costs focuses on time units instead of data units, and attempts to simulate costs of equal length interviews. Reducing the length of the personal interview questionnaire to .6 of its actual size (50 minutes to 30 minutes) would reduce costs of materials preparation (Ann Arbor field office work, typing, duplicating, printing), interviewer salaries and travel for pretest and the final interviewing, and other costs. But with a 30-minute personal interview it is doubtful that costs in any of these areas would be reduced to .6 of their present size. If we merely delete interviewer costs for twenty minutes of questioning, only about $1,700 is saved. But even if all preparation, field, and travel costs (categories II, IV, VI-VII in the table) were reduced by forty percent, the cost of the telephone interview survey would be only 64 percent of that of the personal survey.

We have presented three estimates of the relationship between telephone with personal interview costs. Using unadjusted project figures, sampling and field costs of the telephone survey were about 45 percent of those of the personal, per unit data costs were 50 percent of those in person, and per unit time probably somewhat less than 64 percent of those in the personal interviews.

9. Conclusions

This paper presented findings from an initial study comparing telephone and personal interview surveys. Some of the findings have been replicated by later studies; for example, we continue to achieve lower response rates in national telephone surveys on randomly generated sample numbers than in similar personal interview surveys. Other results may have arisen from our inexperience in administering such telephone surveys; the missing data rate on a series of questions has declined over repeated use of them. Still other results have become inapplicable because of new methodological developments; for example, new sample designs have increased the productivity of telephone interviewers and some costs have changed.

Future work can profitably concentrate on two different areas, 1) interviewer behavior that minimizes response and nonresponse errors, and 2) measurement of nonsampling errors. The identification of optimal telephone interviewer behavior has not yet been achieved; in this project we merely applied techniques found useful in personal interview surveys. However, new interviewer techniques may be desirable for telephone work. The first few moments of telephone interaction where many refusals occur, must form the analogue of a prestudy letter to respondents, the respondent's visual inspection of the interviewer and her written credentials, and all the accompanying descriptive stimuli that a personal interviewer provides a respondent. Now we are merely using trial and error methods in hopes of finding effective introductory techniques, but formal experimental work is required. We have noted that the tendencies toward fast pace in telephone interviews may be associated with more superficial responses to open-ended items. Response effects from questionning speed and interviewer prompting and probing should be formally studied.

All of these suggestions require a data collection design which permits measurement of interviewer effects. Telephone surveys with centralized interviewing staffs permit this more easily than personal interview surveys, and developments in using computer terminals to provide the survey questions to the interviewer and accept the answers of respondents imply that further measures of interviewer behavior may soon be possible. Measurability of these nonsampling errors both aids the evaluation of changes in interviewer behavior and provides the data analyst with better empirical estimates of error in the survey data.

FOOTNOTES

1. On a later survey the status of unanswered numbers was determined and about 95% of the numvers called at least twelve times were not working household numbers. Such unanswered numbers are disproportionately located in rural exchanges where lack of nonworking number recordings is most prevalent.

2. To eliminate one source of differences between modes, we compare telephone survey respondents with personal interview respondents whose households are telephone subscribers.

3. All variance calculations used the ratio mean formula; for the stratified random telephone sample, with elements as ultimate clusters; for the two clustered samples with primary areas as clusters.

4. Because the personal interview sample is larger than the clustered telephone sample, we would expect higher design effects for the personal interview sample. The increase is merely a function of the size of the clusters not of any differences in the sample design, and for that reason we created deft's for an "adjusted" personal interview sample. These figures are presented in the fourth column of the deft's section in Table 5. These were calculated using a sample size of 865, the maximum sample size for the clustered telephone sample.

5. Two estimates, those concerning the respondent's attitude about his life as a whole are measured on half samples. This artifically reduces their design effects for the two clustered samples.

6. We should note that as with most clustered samples, the effects of clustering on the precision of estimates is reduced for analysis of subclasses. For such analyses the clustered telephone sample is relatively more attractive.

7. ρ is a true intraclass correlation coefficient if b is a constant, or does not vary greatly over interviewers. The coefficient of variation of b in the telephone survey was about .09, and we view the presented ρ's as synthetic measures of intracluster homogeneity that also include some effects of varying interviewer load.

8. ρ^*_{int} values were estimated from a $deff_{int}$ using a clustered variance formula with unweighted data. Clusters in the calculations were all interviews completed by a single interviewer; no stratification of clusters was introduced into the calculations.

9. An overall design effect including both sampling design and interviewer effects is approximately

$$Deff_{overall} = Deff_{sampling} + (b_{int} - 1)\ \rho^*_{int}$$

following Hansen, Hurwitz, and Madow's model (1953, Vol. II, pp. 291-293).

10. A group of questions appearing at the end of the personal interview was dropped from the telephone survey questionnaire.

11. If broken-off interviews are included, the total telephone sample size is 1,734.

12. These figures were estimated by hand calculation of number of non-missing data cases in all question sets. Open-ended variables yield two data fields (first- and second-mentioned answers) and were counted as two variables. The figures are so close to one another chiefly because of the larger sample size in the telephone survey.

Selected References

Colombotos, J., "The Effects of Personal vs. Telephone Interviews on Socially Acceptable Responses," Public Opinion Quarterly, XXIX (Summer, 1965), 457-458.

Coombs, L., and Freedman, R., "Use of Telephone Interviews in a Longitudinal Fertility Study," Public Opinion Quarterly, XXVIII (Spring, 1964), 112-117.

Dohrenwend, B.S., Colombotos, J. and Dohrenwend, B.P., "Social Distance and Interviewer Effects," Public Opinion Quarterly, XXXII (1968), 410-422.

Freeman, J., and Butler, E.W., "Some Sources of Interviewer Variance in Surveys," Public Opinion Quarterly, XL (Spring, 1976), 79-91.

Groves, R.M., "On the Mode of Administration of a Questionnaire and Responses to Open-Ended Items," paper presented at MAPOR, 1976.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G., Sample Survey Methods and Theory, II, New York: John Wiley and Sons, Inc., 1953.

Hanson, R.H., and Marks, E.S., "Influence of the Interviewer on the Accuracy of Survey Results," Journal of the American Statistical Association, LIII (1958), 635-655.

Hochstim, J.R., "A Critical Comparison of Three Strategies of Collecting Data from Households," Journal of the American Statistical Association, LXII (September, 1967), 976-989.

Ibsen, C.A., and Ballweg, J.A., "Telephone Interviews in Social Research: Some Methodological Considerations," Quality and Quantity, VII (1974), 181-192.

Kegeles, S.S.; Fink, C.F.; and Kirscht, J.P., "Interviewing a National Sample by Long Distance Telephone," Public Opinion Quarterly, XXXIII (1969), 412-419.

Kish, L., Hess, I., "On Noncoverage of Sample Dwellings," Journal of the American Statistical Association, LIII (June, 1958), 509-524.

Kish, L., "Studies of Interviewer Variance for Attitudinal Variables," Journal of the American Statistical Association, 57 (March, 1962), 92-115.

Klecka, W.R., "Potential Coverage Problems in Telephone Surveys," (Unpublished, 1976).

Rogers, T.F., "Interviews by Telephone and in Person: Quality or Responses and Field Performance," Public Opinion Quarterly, XL (Spring, 1976), 51-65.

Tuchfarber, A.J., and Klecka, W.R., Random-Digit Dialing: Lowering the Cost of Victimization Surveys, Police Foundation, 1976.

Work supported by NSF SOC (76-07519). Cost analysis done in collaboration with Barbara Thomas, of the Survey Research Center Omnibus Survey staff.

DISCUSSION

Harold Nisselson, U. S. Bureau of the Census

Dr. Groves is to be congratulated for a carefully designed and executed study. There are a number of interesting points made in the paper, and questions suggested, and I will comment on a few.

First, the paper again confirms the possibility of serious biases in coverage of the population through a telephone frame. Roughly 1 in 10 households overall are estimated to not have a telephone in the household. The rate estimated from the study is about 1 in 20 for white households and 1 in 6 for nonwhite households, 1 in 6 for households outside SMSA's, and 1 in 6 for those with 1974 income under $7,500. As Dr. Grove points out, for many surveys both the overall coverage and, especially, the differential coverage of subgroups in the population would not be acceptable.

These coverage rates may be viewed as measuring essentially coverage of households. However, coverage of persons within household is at least as important and, in the Census Bureau's experience, more troublesome. Research is needed to assess the extent to which the household is properly defined and acceptable coverage of all persons within households is achieved.

The response rate analysis, as has been noted in other studies, is troublesome -- primarily because of problems in measuring the denominator of the rate. I also would question the average number of calls per household as a measure of effort, since this can be an artifact of the strategy adopted. In some testing of computer-assisted telephone interviewing by the Census Bureau we have had higher telephone response rates than found in the study. This may be due to the auspices. Thus, the paper suggests familiarity with the research organization as an explanation of the lower State-wide rate in Michigan compared to that in the area closest to the telephone interviewing staff. However, it may be that with more experience higher response rates could be achieved. The Census Bureau has not used telephone as the mode for the interviewing of a household for the first time, but we have used it as a supplement to reduce noninterview rates for respondents hard to find at home. Also, in panel studies we make use of telephone interviewing on second and later occasions. We have not found in this context that older people are more easily accessed by telephone as reported by Dr. Groves.

I found the use of the "unfolding measure" in telephoning as a substitute for a flash card in personal interviewing interesting, as are the findings of interviewer influence through the pace of the interview. With regard to the analysis of respondent preferences as to mode, some caution as to the findings may be advisable. The Census Bureau, as I noted, uses telephone interviewing in panel studies on second and later occasions if the respondent when asked is willing to accept it. Interestingly, when we set targets for reducing interviewer mileage in the fuel crises of 1974, the proportion of households in eligible panels that were interviewed by telephone rose substantially. With regard to the question of whether the quality of data obtained by telephone is lower or higher than with personal interviewing, our experience may be summarized as a Scotch verdict. The Census Bureau is planning to carry out extensive controlled studies of this question.

The analysis of sampling and interviewer design effects is interesting, although difficult to follow since the estimators of the various quantities are not given. It appears that in the analysis in Table 5, inadequate account was taken of sample size variation. The large difference in design effects between stratified and clustered telephone interviewing is interesting, but perhaps not surprising. From the point of view of planning a multi-purpose survey, using some quantile of the distribution of design effects over items may be a useful alternative to the average design effect. Any given quantile of the distribution indicates the items and proportion of items which would be subject to design effects no larger than the quantile-value, and hence the proportion subject to greater effects. Viewed this way, there is much less difference -- for example, at the 80-percent point -- between the personal and clustered telephone design effects.

The cost analysis is to be commended, although individual cost factors may differ substantially among organizations.

It is easy to agree with Dr. Groves' conclusions as to the research needs, and to urge his model of controlled experiment. Telephone interviewing is here, and in combination with computers is a much more flexible and potentially useful tool than ever. Now the need is to establish a sound scientific base of knowledge for its use .

DISCUSSION OF "AN EXPERIMENTAL COMPARISON OF NATIONAL TELEPHONE AND PERSONAL INTERVIEW SURVEYS"

Charles D. Palit, University of Wisconsin

By way of an introductory remark, I must say that I am very happy to see Bob Groves doing research on survey methodology and wish to congratulate him on his work. The topic of his paper is indeed an important one to the profession and industry.

Use of the telephone as a data collection instrument provides us with a quantum jump in productivity in terms of cost per bit of information collected. Consequently a knowledge of what else we might be gaining or losing by chosing telephone over personal interviews is important.

Ideally we would like data collected by phone interview to be better than data collected by the personal interview. But, in fact, even if we could conclude that telephone data is "just as good" as personal, it would be cause for a celebration.

From this report, I see that we are not this fortunate with respect to national surveys, for the message of this report is that at this time, we can not make such a clear-cut judgment. We hear that (i) telephone surveys cost less per interview and (ii) tend to produce a smaller sampling error than personal interview surveys--this is as expected--and (iii) that response differences between the two modes are minimal, which is fortunate or else we might be stuck with trying to decide which was more accurate.

The two response differences detected can readily be ascribed to our inexperience with the telephone mode. I like to think that by working on it, we can increase the satisfaction of the respondent with the telephone interview.

Let us look at the first item with a response difference: "The Frequency of Missing Data." Here Bob reports a higher incidence with telephone but also reports that this problem declined as the interviewers gained experience. The Wisconsin Survey Research Laboratory's experience is that a centralized phone operation allows for much closer supervision of interviewers and an earlier correction of procedural errors. Further, the low cost of a verification call allows us to routinely make post-interview calls on the respondent, as part of our editing process.

Turning to the second item, which I will label "fewer responses to open-end questions," the example cited -- 11 percent fewer phone respondents supplied three or more problems facing the country. Even though the probing on this question was well controlled, my suspicion is that timing is a problem. I suspect that in the absence of visual cues, the interviewer did not allow as much time for the respondent to respond on the phone as was done in the personal interview situation. As further support for this hypothesis, I note that this item has the highest interviewer intra-class correlation coefficient reported, approximately .07; indicating perhaps a higher than average sensitivity to interviewer effect. More training on the timing of probes may well eliminate the 11 percent difference in response frequency.

In addition to these response differences, there are, of course, pieces of observational information which can not be recorded by the interviewer using the telephone mode, or by the sampler. A good example is the size of the place in which the respondent's housing unit is located. We can query the respondent for this information, but the information provided is likely to be less accurate than the observational information provided by the personal interview mode. In fact, as with anything else, the question used to gain this information will influence the quality of the information obtained.

Table 1 is a good illustration of this. In one Wisconsin telephone survey, we asked each respondent two questions, the first as to the approximate size of the population in their minor civil division (MCD) of residence, and the second as to the name of the MCD. Later the population size corresponding to the MCD named was coded. Table 1 shows the percent agreement between population size which resulted. Overall, approximately 20 percent of the responses disagreed.

TABLE 1

PERCENT AGREEMENT ON TWO METHODS OF DETERMINING POPULATION SIZE OF PLACE OF RESIDENCE BY REPORTED POPULATION SIZE

Reported Population Size of Residence	Percent Agreement
Less than 2,500	81
2,500 - 9,999	60
10,000 - 24,999	74
25,000 - 49,999	76
50,000 - 99,999	88
100,000 or over	93
Not ascertained	23

Now what about...(i) the population coverage provided by the sample--a combination of coverage provided by the frame and response rate, and (ii) what Bob has called $Deff_{int}$ i.e., the interviewer effect contribution to the variance?

With regard to the coverage problem, it is important to emphasize that Bob's results of a 90 to 93 percent frame coverage and 59 to 70 percent response rate can be improved when we are dealing with smaller areas. Bob has already pointed out that the closer to home, the better the response rate. This is consistent with our experience in Wisconsin, but in addition some states have better frame coverage than others. For example, Wisconsin's telephone frame coverage as estimated by personal interview survey is about 95 percent. This, with a response rate of say 80 percent, would give us an overall coverage rate of 76 per-

cent, so that at least for some areas, we can begin to get close to the overall coverage rates usually achieved by the personal interview mode.

But even for the national survey, the situation is a bit better than painted if our population of interest is adults residing in housing units. For example, from Bob's data on the number of adults in non-phone housing units, we can easily see that because non-phone housing units have fewer adults, the frame coverage rates for the adult population move up about one percentage point. This may seem small, but if we consider its value in terms of what it would cost to raise the response rate one percentage point, it is a handsome gift.

I think that the most disturbing part of Bob's report for me was the discussion of the contribution of interviewer effect to the variance of our estimates as measured by $Deff_{int}$.

The nature of the telephone operation is such that a substantially greater proportion of the interviewer's time is spent on interviewing than is the case for personal interviews; consequently the number of interviews produced by each interviewer is much larger on the average for telephone than for the personal mode. What is disturbing is that even though better control of the interviewers in a centralized operation may lead to smaller interviewer intra-class correlations, the larger number of interviews per interviewer will tend to inflate the $Deff_{int}$.

If we want to reduce this, we have the choice of finding better methods of controlling interviewer effect or reducing the interviewer's work time. If we reduce the interviewer's work time too much, then it may not be worth the interviewer's time to work nor our time to train them. Of course, this may still be preferable to the confounding of the interviewer effect with location that takes place in the usual area probability sample.

In conclusion, I must say that I believe we have only scratched the surface in our development of telephone survey methodology, and we can expect further improvements to be forthcoming which will make this mode even more competitive with the personal mode. More methodological studies are necessary for this. They cost money, but in terms of what they will do for the productivity of the social sciences, I think it would be money well spent.

* * * * *

PREDICTING THE CHANGE IN THE NUMBER OF PERSONS IN POVERTY BY A REGRESSION MODEL 1/

Bruce W. Klein, U. S. Department of Health, Education, and Welfare

Introduction

The number of persons in poverty measured by the CPS series is arrived at by comparing incomes of families and unrelated individuals from the annual March supplement on the Current Population Survey to Orshansky Poverty Thresholds. Those families and unrelated individuals falling below the thresholds are considered poor, those falling above the thresholds are nonpoor. 2/ Any references to poverty that follow refer to this official measure.

Due to the large increase in poverty in 1975 the idea occurred to us to investigate the relation between the CPS poverty series and exogenously determined macroeconomic variables. Year-to-year percent changes in real GNP (Gross National Product) and the unemployment rate were thought to be the best theoretical predictors of changes in the number of poor. 3/ A regression model yielded an R^2 of .88 with highly sifnificant coefficients bearing out the implicit hypothesis that year-to-year changes in the poverty series reflect year-to-year changes in aggregate economic performance.

The Model

The general form of the equation is:

(1) Number of persons in poverty = f(real GNP, Unemployment rate)

Real GNP is an indicator of economic performance while the unemployment rate is a measure of the economy's utilization of experienced workers. It is well known that the GNP growth rate is an indicator of changes in the minimum standard of living. 4/ When the economy expands real GNP rises. As this process occurs employed workers and the marginally employable make a larger contribution to output. As the intensity of the contribution of these workers increases their incomes increase. It is thought that many of these workers come from low income families that fall in and out of poverty due to the contribution these workers make to their incomes. When a families' standard of living rises, they come out of poverty; when the standard falls they go into poverty. Approximately 60% of the poor had at least one family member that worked in each survey year. A higher percent of families with at least one worker is found among families that are below 125% of the poverty level. 5/

As the economy expands the unemployment rate also decreases. The affect of changes in the unemployment rate on poverty is of smaller consequence when compared to the affect of changes in real GNP on poverty. The affect is smaller because only about 9% of heads of poverty families are officially unemployed.

So we have isolated a poverty effect due to a change in aggregate economic performance. When GNP and employment go up poverty goes down. A very intuitive Keynesian result. The factor linking the two is the increased contribution of workers at the marigin who would fall in units below the poverty threshold without the rise in economic production. The effect also occurs in reverse when GNP and employment go down.

Maybe the poverty status of persons not able to work, the aged, disabled, and female heads with very young children, have possibly been constant or slowly lessened over time and therefore do not attribute much variation to year-to-year changes. Their income is dependent upon transfer payments which have a more complicated relation to economic performance. The effects on non-working poor of changes in economic performance should be the subject of another paper.

The specific form of the model is:

(2) $\#POOR = C + B\ GNP + B\ UNEMP$

where

C = constant

#POOR = percent change in the number of poor
GNP = percent change in Gross National Product in 1972 constant dollars (real GNP growth rate)
UNEMP = Annual official unemployment rate

Each year from 1959-1975 accounts for one observation. So the model for all 16 years produces a final form of the ordinary least squares regression equation:

(3) $\#POOR = -5.8443 - 1.4651\ GNP + 1.6724\ UNEMP$
(2.9) (.23) (.47)

Below the coefficients in parenthesis appear the standard errors. All coefficients are significantly different from zero at the 95 percent confidence level (within two standard errors). Table 1 is a table of standard errors, t-values, and analysis of variance. Coefficients are tested against the null hypothesis that the coefficient equals 0. Table 2 is a table of the actual values of the independent variable, estimated values, and residuals.

TABLE 1

Variable	Coefficient	Std. Error	t -value	Significance at 99% level
c(constant)	-5.8443	2.9079	-2.0098	*NS
GNP	-1.4651	.2284	-6.4136	S
UNEMP	1.6724	.4694	3.5631	S

		Analysis of Variance	DF	Sum of Squares	Mean Square	F-test	Significance at 99% level
Multiple R	.9385						
R^2	.8807						
Adjusted R^2	.8624	Regression	2	404.24	202.12	47.997	S
Std. Error	2.052	Residual	13	54.743	4.2110		

NS = not significant
S = significant
* significant at the 90% level

TABLE 2

Percent Change in the Number of Persons in Poverty

Period	Actual Percentage Change	Estimated Percentage Change	Residual
1959-60	1.0	- .02	1.02
1960-61	- .6	1.70	-2.30
1961-62	-2.5	-5.14	2.64
1962-63	-5.7	-2.17	-3.53
1963-64	-1.1	-4.91	3.81
1964-65	-8.0	-6.96	-1.04
1965-66	-9.1	-8.28	- .82
1966-67	-2.6	-3.44	.84
1967-68	-8.6	-6.27	-2.33
1968-69	-4.3	-3.80	- .50
1969-70	5.1	2.79	2.31
1970-71	.3	- .37	.67
1971-72	-4.5	-4.83	.33
1972-73	-6.1	-5.71	- .39
1973-74	5.6	6.01	- .41
1974-75	10.7	11.01	- .31

Limitations of the Model

Two known sources contribute to the model's limitations. Both sources are due to the nature of the CPS survey data. In the first case, the C. V.[6] dropped steadily from 1.85% in 1959 to 1.32% between 1966-75. The marked change in the C.V. was due to an expansion of the sample in 1967 (1966 data) from 33,000 in 1966 to 48,000 households in 1967. The sample became 45,000 households in 1971. Thus, the standard error varies from one year to the next.

The second source is a function of the sample selected for the survey. Year-to-year overlap in the sample affects the variation in the number of poor persons estimated by the model. In the Current Population Survey (CPS), there are eight rotation groups. The groups are in the sample for four months out of the sample for eight months and back in the sample for four months in rotating order. A 50% overlap in the sample of households results. There are not necessarily a sample of the same household occupants, but 50% of the same addresses are sampled from one year to the next for each given month. A year-to-year correlation coefficient for poverty estimates results as shown below:

Years	Persons	Families
1974-1975	0.40	0.35
1971-1972	0.15	0.14
1970-1971	0.31	0.28

The positive year-to-year correlations reduce the variance of the number of poor persons estimated by the model.

Current Estimates

By using the model as a point predictor, an estimate for 1976 can be computed as an illustration. [7]

The values for the variable for 1976 are:

GNP = 6.1%
UNEMP = 7.7%
Substituting these into the equation yields:

#POOR = -1.904(%)

By multiplying and then adding that result to the number of poor in 1975 a 1976 estimate of the change in the number of poor and an estimate of the number of poor can be derived.

Chg. in the no. poor = #POOR X Actual no. of poor in 1975

= -1.904 X 25,877,000

= -492,000

Est. no. of poor 1976 = Chg. in the no. poor + Actual no. of poor

= -492,000 + 25,877,000

= 25,385,000

By using the standard error (.0205) a 95% confidence interval can be constructed around the estimates yielding:

29,000 CHG. in the no. poor -1,012,000

24,345,000 EST. no. of poor 1976 26,426,000[8]

Conclusion

The CPS poverty series follows along well in year-to-year changes with variables that measure the macroeconomic performance of the economy. This relationship bears out the well known statement that GNP growth has provided absolute increases in the U. S. minimum standard of living as evidenced through the poverty thresholds. The findings of this paper also give support to the meaningfulness of CPS income data and the Orchansky poverty measure in light of recent criticism of both. [9]

FOOTNOTES

1/ Thanks goes to Renee H. Miller who assisted in the statistical methodology and interpretation of results, but alas, all responsibility for the final draft goes to the author.

2/ See Current Population Reports, Series P-60, No. 102, Appendix A.

3/ After the research was completed, it was learned that the percent point change was used in Okun's work on the relationship between GNP and unemployment. Only further research can determine if the point change in the unemployment rate is a better predictor of poverty than the unemployment rate. For more information refer to Arthur Okun's, The Political Economy of Prosperity, W. W. Norton and Co.,New York, 1970.

4/ P. A. Samuelson, Economics, (McGraw-Hill, New York, 1973, 9th ed.), p. 80.

5/ Current Population Survey, U. S. Bureau of the Census.

6/C. V. is the coefficient of variation on the estimated number of persons in poverty. It is defined to be the standard error of the estimate divided by the estimate.

7/ A more current estimate is not available since the Bureau of the Census has not yet released 1976 actual data.

8/ All numbers are rounded to the nearest thousand to conform with Bureau of the Census convention.

9/ See The Measure of Poverty, U. S. Department of Health, Education and Welfare, April 1976;

and Poverty Status of Families Under Alternative Definitions of Income, Background Paper No. 17, Congress of the United States, Congressional Budget Office, Washington, D. C., January 13, 1977.

REFERENCES

Bureau of the Census. Appendix A, Current Population Reports, Series p. 60, No. 102, January 1976.

Johnston, J. Econometric Methods. McGraw-Hill Book Company, Inc. New York, 1963.

Merrill, William, C. and Karl A. Fox. Introduction to Economic Statistics. John Wiley and Sons, Inc., New York, 1970.

Okun, Arthur. The Political Economy of Prosperity. W. W. Norton and Co., New York, 1970.

Samuelson, Paul A. Economics. McGraw-Hill, New York, 9th Edition, 1973.

MEASURING THE SOCIOECONOMIC STATUS OF OCCUPATIONS

Alice Henry, Cornell University
Neil W. Henry, Virginia Commonwealth University

During the past 10 years a great deal of systematic sociological analysis has been based on the socioeconomic index (SEI) developed by O.D. Duncan and his associates (Duncan, 1961). This index is a simple function of the income and education distributions within an occupational category and as such can be computed from available census data for relatively narrow occupational classifications. Replacing earlier scales of socioeconomic status which were based on attributes which were either very difficult to measure or which reflected ad hoc decisions of an individual researcher, the SEI has enabled sociologists to cumulate knowledge of occupational attainment and mobility from one study to another.

Duncan's SEI was calculated from the distribution of income and education of males in each detailed census category in 1950. The specific equation adds together .59 times the percentage of men with at least four years of high school (Blau and Duncan, 1968: 125). (There is also a constant added, which is irrelevant to our discussion.) The SEI was validated and the coefficients mentioned above determined by regressing occupational prestige measured in studies conducted by the National Opinion Research Center (NORC) on the two predictor variables. The NORC scores were available only for a limited number of occupations, and Blau and Duncan report an R^2 of .83 using data on 45 occupations.

The NORC scale is based on responses of the public-at-large to questions such as: "Which statement on this card best gives YOUR OWN OPINION OF THE GENERAL STANDING OF A RAILROAD BRAKEMEN? What number on that card would you pick out for him?" (Reiss, et.al., 1961, Appendix A, their caps). The instructions clearly refer to men. To eliminate any remaining chance that the index could be applied to women, the designers of the NORC study deliberately left out "women's occupations":

> To keep the number of occupations within the practical limits of the NORC study, this original list of 100 occupations was reduced to 78, primarily by eliminating "women's occupations", such as private secretary, dress maker, trained nurse, and domestic workers, and others thought to be already covered by the continuum. Parenthetically, it might be noted that some of these deletions in the interest of practicality appear to have impaired the "representativeness of the list." (Reiss, 1961:5)

Nevertheless these indices <u>have</u> been used to study the occupational status and mobility of women, and to draw conclusions about the relative status of men and women, and of men's and women's occupations (Treiman and Terrell, 1975; McClendon: 1976). The purpose of this paper is to show that such application of a male-based index to the female or the entire labor force is improper, and cannot help but lead to misleading results when used to compare women's occupational status to that of men.

In order to correct the unrepresentativeness of the NORC/SEI procedure one would have to study the general prestige of a list of occupations that included "women's occupations", and explicity use women as well as men as referents when describing the jobs. (An alternative methodology would ask people to rate separately the standing of male and female occupants of the same job.) In her dissertation Bose (1973) conducted such a study, but did not take the next step. that is. to use the income and educational attainment of all persons in the labor force to calculate an index for each detailed occupational category. Such an index would be validated and optimal coefficients determined as in the case of the traditional SEI devised by Duncan. We have not carried out such a study: rather, by using the fundamental idea of the Duncan SEI, we have merely carried out an exercise to verify that the SEI based on the male labor force <u>does</u> misclassify women workers and "women's occupations", and that the conventional male-based index is not as adequate as Parnes (1970), Treiman and Terrell (1975) and McClendon (1976) have implied.

Using 1970 census data two SEI scores were calculated for the 588 detailed occupational categories: the traditional one based on male occupants only and the other based on all occupants of the category. In both cases the same index was used, namely

$$\text{SEI} = .5\,(\%\text{ with income over \$8000}) + .5\,(\%\text{ with at least one year of college}).$$

The procedures closely paralleled those used by Duncan: the cutting points in the income and education distributions are at approximately the same percentiles as the 1950 figures; the entire experienced worker labor force is used, rather than full-time workers; the

weights used by Duncan are nearly equal. When we compared our male-based scores to Duncan's 1950 (male-based) scores, we found little difference in the relative standing of the major occupational groups.

Severe discrepancies appear, however, when the relative standings of some occupations are compared on the different sets of scores. For example, the title "secretaries" includes 2,770,426 workers, 98% of whom are female. On the male-based scale this occupation is 9 points above the mean, while when the scale based on all workers is used we find that secretaries are 14 points below the mean for all occupations. In table 1 we have summarized comparisions of this type, considering an occupation to be classified differently by the two scales whenever there is more than five points difference in the scores, relative to the respective means. (e.g., for secretaries this difference would be 23 points.) Using this criterion 103, or 18%, of the 588 occupations are classified differently by the male-based and all-person-based scales. These occupations, moreover, contain 46% of all the women in the labor force. Discrepancies are most noticable in a major occupational grouping like "clerical", where 33 of 50 detailed occupational categories are classified differently; the 33 occupations contain 92% of all the female clerical workers.

While the ranking given to the 588 occupational categories by the men-only scale is highly correlated with the all-person scale, this correlation masks the fact that a substantial number of occupations are ranked differently. More importantly, the fact that the male scale misstates the status of so many women casts doubt on the claim of Treiman and Terrell 1975:182) that:

> "it is clear that labor market discrimination against women does not extend to the status of the work open to them nor to the qualifications demanded. Women work at jobs which are about as prestigious as those held by men and, like men, secure good jobs mainly on the basis of superior education."

Table 2 shows the distributions of occupational status of men and women that we found when the scale based on the entire labor force was used as the measure of status. The median status of women is some 9 points lower than that of men: 15 vs. 24. The clustering of women in low status occupations is particularly apparent. 72% of women work at jobs with status scores below 20, compared with only 36% of men. These results support the hypothesis that women are, in fact, excluded from relatively high status occupations.

The exercise reported here confirms our intuitive feeling that socioeconomic indices of occupations based on male data should not be used to evaluate the occupational attainment of women or to compare their attainment to that of men. Any future work applying the status attainment model of Blau and Duncan to women must use a scale of occupational status that is based on both men and women. Theoretically, there is no justification for excluding the female labor force from consideration when estimating the socioeconomic status of an occupation. Methodologically, it leads to serious error.

References

Blau, Peter M. and Otis Dudley Duncan, 1967, The American Occupational Structure. New York: Wiley.

Bose, Christine E., 1973, Jobs and Gender: Sex and Occupational Prestige. Baltimore: John Hopkins University, Center for Metropolitan Planning and Research.

Duncan, Otis Dudley, 1961, "A socioeconomic index for all occupations," in Reiss, et. al., 1961.

McClendon, McKee J., 1976, "The occupational status attainment processes of males and females." American Sociological Review 41 (February): 52-64.

Parnes, Herbert S., John R. Shea, Ruth S. Spitz and Frederick A. Zeller, 1970, Dual Chareers. Vol. I. Washington D.C.: U.S. Department of Labor, Manpower Research Monograph No. 21.

Reiss, Albert J., et. al., 1961, Occupations and Social Status. Glencoe: Free Press.

Treiman, David and Terrell, Kermit, 1975, "Sex and the process of status attainment: a comparison of working women and men." American Sociological Review, 40(2): 174-201.

U.S. Bureau of the Census, U.S. Census of the Population: 1970. Subject Reports. Occupational Characteristics. Final Report PC(2) 7A. Washington D.C.: U.S. Govt. Printing Office.

Table 1: The number and proportion of occupations and women in them that are classified differently when measures of occupational SES are based on all persons rather than on only men, by major occupational group.*

Major Occupational Group	Occupations			Female Labor Force		
	N	%	Total	N	%	Total
Professional, technical and kindred workers	19	15%	127	1,787,449	38%	4,674,716
Managers and Administrators	11	17%	63	419,868	39%	1,083,601
Sales	9	53%	17	928,531	41%	2,249,259
Clerical	33	66%	50	9,724,953	92%	10,515,431
Craftspersons	4	4%	92	96,998	18%	547,761
Operatives	12	11%	114	591,588	13%	4,430,853
Transport	0	0%	12	0	0%	138,979
Laborers	1	2%	61	4,498	1%	307,688
Farm and farm laborers	0	0%	8	0	0%	253,558
Service	12	32%	38	304,219	21%	5,061,341
Private Household	2	33%	6	249,137	21%	1,186,369
Total	103	18%	588	14,107,217	46%	30,534,658

Table 2: Distribution of occupational status by sex, using the measure of socioeconomic status of occupation based on the entire experienced civilian labor force.*

SES of occupation	Men	Women
0-4	1.31%	10.31%
5-9	11.24	19.07
10-14	9.04	18.28
15-19	14.15	23.95
20-24	15.44	5.03
25-29	5.37	0.98
30-34	6.48	2.10
35-39	5.52	3.98
40-44	4.16	1.24
45-49	4.18	1.97
50-54	4.33	2.64
55-59	2.77	1.09
60-64	3.95	1.96
65-69	3.49	4.55
70-74	2.68	1.72
75-79	0.89	0.39
80-84	2.84	0.40
85-89	0.90	0.20
90-94	1.37	0.14
	100.01% (49,518,235)	100.00% (30,449,555)

* Source: U.S. Bureau of the Census. U.S. Census of the Population: 1970. Subject Reports. Occupational Characteristics. Final Report PC(2) 7A. Washington D.C.: U.S. Govt. Printing Office.

ESTIMATION OF SUMMARY MEASURES OF INCOME SIZE DISTRIBUTION FROM GROUPED DATA

Emmett Spiers, U.S. Bureau of the Census

The impetus for this research springs from a decision by the Census Bureau to update its historical series on the trends in the income of families and persons [1]. Since such an undertaking involved the handling of truly massive amounts of grouped income data, it was necessary to employ methods for calculating summary distribution measures which were inexpensive as well as reasonably accurate. In the present paper we will discuss the methods finally chosen and compare them to some of the alternatives considered.

Organizationally,the paper is divided into four sections. The first of these provides a brief overview of available techniques and describes the properties we will require for our application. Sections 2 and 3 discuss some numerical comparisons made between various alternative estimation procedures. Section 4 provides a few concluding remarks.

1. PROPERTIES DESIRED AND ALTERNATIVES CONSIDERED

As a preliminary to the work discussed in this paper,a number of desired properties were set down as requirements. There were four general criteria imposed:

(1) The method should fit the given points exactly (no curve fitting).

(2) Some bias in the estimates can be allowed providing it is consistent; i.e., the estimation technique should not introduce spurious trends into the data.

(3) Simple and efficient methods are best, if possible.

(4) All the summary measures from the grouped data (quantiles, income shares, Gini ratios, etc.) should be consistent with one another and with income distribution theory (i.e., the distribution functions and Lorenz curves obtained should always be nondecreasing).

Since the entire historical series to be updated comes from the Current Population Survey (CPS), several more criteria were imposed that were tailored specifically to that survey:

(5) The data should be"smoothed"somewhat to allow for rounding in the CPS [2].

(6) Because of the widths of the upper income intervals, methods consistent with the theory of income distribution [e.g., 3] are preferable.

(7) The method should be able to handle unusually shaped income distributions; e.g., the method will be used for doctors and surgeons, as well as for service workers.

(8) Since the mean incomes per income interval generally are unavailable for the major portion of the series, the method has to be one which does not depend on this information.

Some of the best known interpolation procedures for income data are precluded by these requirements. In particular, the techniques suggested by Gastwirth-Glauberman [4] and Budd [5] both employ knowledge of the mean income in each interval. A number of general purpose interpolation techniques, unless modified, also lack one or more of the above properties. Two, for instance, that we examined and which proved unsatisfactory were cubic spline interpolation [6] and Akima's method of Local Procedures [7,8]. 1/

From a companion paper by Oh [9] we did have available a general purpose interpolation scheme, Karup-King osculatory interpolation, which had been modified to handle income data. 2/ In the next section we will compare Oh's procedure with the combination of Pareto and linear interpolation we suggest here. The Hermite interpolation technique advocated by Gastwirth-Glauberman will also be considered, even though it cannot always be used in the CPS.

2. ESTIMATING INCOME QUANTILES IN THE CPS

In this section we will examine three different methods for estimating income quantiles from the CPS. The three methods are--

(1) Actual quantiles--The "actual" quantiles from the ungrouped CPS data were calculated by sorting the CPS microdata files and picking the income representing each of the quantiles selected for comparison (i.e., the 20th, 60th, 80th and 95th percentiles). This was done separately for families (table 1) and unrelated individuals (table 2) for each income year 1958-1974.

(2) Pareto-linear--The Pareto-linear estimates were developed assuming uniform distributions in the lower income intervals and Pareto distributions in the upper income intervals. The starting point of the calculations was annual Census Bureau CPS income reports. Each interval was interpolated separately. Pareto interpolation was used whenever the absolute value of Pareto's slope parameter was greater than 1. Usually this condition occurred in the income intervals above the median. The absolute value of this parameter is generally greater than 2,in the top interval,and decreases as income decreases. Pareto interpolation could have been used

even after the parameter became less than one; however, we did not use it, because the estimates derived from Pareto interpolation were frequently less accurate than those derived from linear interpolation.

(3) Karup-King osculatory interpolation--The third method used was Karup-King osculatory interpolation, modified as necessary for use with income data [9]. For the comparisons in this paper, we first converted the income and frequency information to a log scale, in order to better graduate the distributions in the longer intervals in the upper tail. Basically, the procedure consisted of deriving the cumulative distribution function in the interval [b, c) by examining the interval just before it, say [a, b), and just after it, say [c, d). Two quadratic equations were then fit through the points {a, b, c} and {b, c, d}. These two quadratic equations were then weighted in such a way as to force a smooth nondecreasing cumulative distribution through b and c. Moreover, the procedure had to fit a, b, c, and d exactly. An extra point was provided in the top open-end interval by fitting a Pareto distribution to the interval preceeding the open interval and estimating the frequency above $100,000.

Now that we have outlined the three methods to be looked at, it is appropriate to turn to the actual (numerical) comparisons in tables 1 and 2.3/ Several observations are possible:

(1) Relatively speaking, income quantiles can be more accurately estimated for families than for unrelated individuals (i.e., both interpolation procedures tend to be relatively closer to the ungrouped data for families than for unrelated individuals).

(2) The pattern of accuracy is also different for families than for unrelated individuals. For families, the data are better for the lower quantiles than for upper quantiles, while the reverse is true for unrelated individuals. Undoubtedly, this pattern occurs for families because the income intervals used to calculate higher quantiles are much broader than for lower quantiles. However, for unrelated individuals, lower quantiles fall in the extreme bottom intervals, where the size of the interval is still large relative to the magnitude of the estimate being attempted.

(3) The CPS data follow the Pareto law rather closely in the upper tail of the income distribution, as has been mentioned, especially if one fits the CPS to a Pareto which can change from interval to interval, as is done here. This is one of the main reasons the Pareto-linear interpolation works so well.

(4) The Pareto-linear procedure seems to provide more accurate measures more of the time than does the Karup-King. This was in some sense unexpected because Karup-King, as employed in this paper, essentially represents a refinement to a simple log-log (Pareto type) interpolation procedure. I suspect that the Karup-King might have been better had we accumulated the data from higher intervals to lower intervals and then applied the osculatory interpolation formulas.

3. LORENZ CURVE ESTIMATION IN THE CPS

We now turn from the interpolation of income quantiles to obtaining selected Lorenz curve measures (income shares and Gini ratios). Again, we will make comparisons (in tables 3 and 4) between three methods:

(1) Actual values--For each year we calculated the aggregate income received by each percentile of the population. This was done separately for families and unrelated individuals from CPS microdata files sorted by amount of income. In table 3 we look at just families over the period 1967-1974 so as to be consistent with [4]. In table 4 we examine the entire time series.4/

(2) Pareto-linear--To obtain Lorenz curve values using this method, the aggregate income in each size class had to be derived. We did this by assuming Pareto distributions in each income interval for the higher intervals and assuming a uniform distribution in the lower intervals. The same decision rule as before was used for switching from one method to the other. In the top open-end interval, the frequency with income above $100,000 was estimated from a Pareto distribution fitted to the previous interval. An assumed mean of $100,000 was assigned to units with income over $100,000. The closed interval form of the Pareto mean income estimation formula was used for the remaining units in the open-end interval (see [10] for full details). The Gini index was estimated by splitting the given Lorenz curve into 100 intervals, each of one percent, and using Simpson's rule for approximate integration.

(3) Hermite--Gastwirth and Glauberman [4] employed Hermite interpolation to develop Lorenz curve measures from the CPS for the years 1967-1974. We have reproduced these here, in part, because Karup-King estimates were not available in time for the presentation at the session.

At least two overall observations seem in order for the comparisons in the tables:

(1) The Hermite interpolation procedures of Gastwirth and Glauberman assume that mean income per income interval is known. For this reason,we expected their estimates to be better than the Pareto-linear ones,since,for the latter,the actual means in each interval are not used. However, the results seem to indicate that the Pareto-linear method is slightly more accurate than Gastwirth-Glauberman's. I suspect, though, that data for all families do not represent an adequate test. It is my opinion that Hermite interpolation might be better than Pareto-linear for unrelated individuals or for race data.

(2) For Gini indexes, the Pareto-linear differs from the ungrouped data by, at most, .004, while Gastwirth-Glauberman differs by, at most, .006. Both methods tend to underestimate the Gini index slightly. However, neither method appears to introduce a spurious trend. Similar closeness to ungrouped data is indicated for shares of aggregate income.

4. CONCLUSIONS

This paper examines several methods for estimating summary measures of income distributions from grouped data. Of those considered in detail, it would seem that the Pareto-linear is best suited for our application to the Current Population Survey historical income series. The advantage of the method grows when one considers its simplicity and ease of use. In fact, the Census Bureau has adopted Pareto interpolation for calculating published CPS medians when these fall in intervals of more than $1,000 in length. This will be fully implemented for the annual 1976, series P-60, income report.

ACKNOWLEDGEMENTS AND FOOTNOTES

The author would like to thank Fritz Scheuren and H. Lock Oh of the Social Security Administration for their many helpful comments and suggestions. Thanks are also due to Professor Gastwirth of the George Washington University for his stimulating discussion of an earlier draft of this paper. Editorial assistance was provided by Mary Henson and Gordon Green. The typing was done by Rubye Ellis.

1/ It is possible that we were not patient enough in applying these methods; even as trivial a modification as converting to logs before interpolating may well have yielded acceptable results. However, given the comparisons made with Karup-King Osculatory Interpolation [9], we suspect that these methods would generally not be better than the simpler (Pareto-linear) technique actually adopted.

2/ Oh's procedure satisfies all our requirements with the exception of perhaps number 6.

3/ Very little work has been done so far to estimate the standard errors of the differences among the several interpolation methods presented. Sampling error is not, however, likely to be a serious limitation on the comparisons in the tables, since each of the methods was applied in turn to exactly the same data sets, the March CPS's from 1959 to 1975 (i.e., income years 1958-1974, respectively).

4/ This paper does not represent the first appearance of these ungrouped figures in print. Most of them were originally prepared by me several years ago and published in Series P-60 beginning with report No. 90.

REFERENCES

[1] U.S. Bureau of the Census, "Trends in the Income of Families and Persons in the U.S., 1947-1964," Technical Paper No. 17.
U.S. Bureau of the Census, "Trends in the Income of Families and Persons, 1947-1959," Technical Paper No. 8.

[2] Knott, J., "An Analysis of the Effect of Income Rounding in the Current Population Survey," 1971 American Statistical Association Proceedings, Social Statistics Section, 1972.

[3] Bjerke, K., "Income and Wage Distributions - Part I: A survey of the literature," Review of Income and Wealth, Series 16, No. 3 (Sept. 1970), pp. 235-252.

[4] Gastwirth, J.L., and Glauberman, M., "The Interpolation of the Lorenz Curve and Gini Index from Grouped Data," Econometrica, Vol. 44, pp. 479-483, May, 1976.

[5] Budd, E.C., "Postwar Changes in the Size Distribution of Income in the U.S.," American Economic Review, 60 (May 1970), pp. 247-260.

[6] Sperry-Rand Corporation, Univac Division, Math-Pack, Programmers Reference, UP7542, 1967, section 2, pp. 68-74.

[7] Akima, H., "A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures," Journal of the Association for Computing Machinery, Vol. 17, No. 4 (October, 1970), pp. 589-603.

[8] Akima, H., "Interpolation and Smooth Curve Fitting Based on Local Procedures," Collected Algorithms from Communications of the Association for Computing Machinery, Algorithm 433, March 1, 1972.

[9] Oh, H.L., "Osculatory Interpolation with a Monotonicity Constraint," 1977 American Statistical Association Proceedings, Statistical Computation Section.

[10] Spiers, E., "Some Notes on the Derivation of Computation Formulas Assuming a Pareto Distribution," (Unpublished Working Paper), 1976.

Table 1.--COMPARISON OF DATA ON SELECTED INCOME QUANTILES FOR FAMILIES BY TOTAL MONEY INCOME IN 1958 TO 1974, BY TYPE OF ESTIMATION METHOD, FOR THE UNITED STATES

Year	Ungrouped Data	Pareto-Linear	Percent Difference (2)-(1)/(1)x100	Karup-King Log-Log Scale	Percent Difference (4)-(1)/(1)x100	Ungrouped Data	Pareto-Linear	Percent Difference (7)-(6)/(6)x100	Karup-King Log-Log Scale	Percent Difference (9)-(6)/(6)x100
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	TWENTIETH PERCENTILE					SIXTIETH PERCENTILE				
1974	$6,500	$6,551	0.8	$6,552	0.8	$14,916	$14,944	0.2	$14,948	0.2
1973	6,081	6,141	1.0	6,143	1.0	14,000	13,883	-0.8	14,012	0.1
1972	5,612	5,668	1.0	5,671	1.1	12,855	12,816	-0.3	12,932	0.6
1971	5,211	5,275	1.2	5,277	1.3	11,826	11,850	0.2	11,873	0.4
1970	5,100	5,148	0.9	5,154	1.1	11,299	11,337	0.3	11,404	0.9
1969	5,000	5,005	0.1	5,005	0.1	10,800	10,799	-	10,883	0.8
1968	4,544	4,598	1.2	4,610	1.5	9,960	9,968	0.1	9,970	0.1
1967	4,097	4,164	1.6	4,172	1.8	9,000	9,129	1.4	9,137	1.5
1966	3,935	3,950	0.4	3,951	0.4	8,563	8,644	0.9	8,664	1.2
1965	3,500	3,508	0.2	3,508	0.2	7,910	7,982	0.9	7,991	1.0
1964	3,250	3,288	1.2	3,288	1.2	7,500	7,574	1.0	7,601	1.3
1963	3,096	3,150	1.7	3,156	1.9	7,134	7,223	1.2	7,244	1.5
1962	3,000	3,018	0.6	3,019	0.6	6,800	6,851	0.8	6,863	0.9
1961	2,800	2,827	1.0	2,831	1.1	6,560	6,631	1.1	6,663	1.6
1960	2,784	2,795	0.4	2,799	0.5	6,364	6,423	0.9	6,451	1.4
1959	2,677	2,715	1.4	2,719	1.6	6,081	6,176	1.6	6,194	1.9
1958	2,530	2,558	1.1	2,556	1.0	5,720	5,774	0.9	5,817	1.7
Average absolute % difference			0.9		1.0			0.7		1.0
Maximum % diff			1.7		1.9			1.4		1.9
Number times better			9		1			14		1
Maximum less minimum			1.6		1.8			2.4		1.8
Year	EIGHTIETH PERCENTILE					NINETY-FIFTH PERCENTILE				
1974	$20,445	$19,894	-2.7	$20,968	2.6	$31,948	$31,957	-	$30,562	-4.3
1973	19,253	18,658	-3.1	19,596	1.8	30,015	30,296	0.9	28,970	-3.5
1972	17,760	17,418	-1.9	18,058	1.7	27,836	28,152	1.1	27,072	-2.7
1971	16,218	16,119	-0.6	16,370	0.9	25,325	25,520	0.8	25,310	-0.1
1970	15,531	15,538	-	15,633	0.7	24,250	24,342	0.4	24,597	1.4
1969	14,751	14,783	0.2	14,815	0.4	22,703	22,757	0.2	23,435	3.2
1968	13,400	13,434	0.3	13,556	1.2	20,590	20,664	0.4	21,168	2.8
1967	12,270	12,395	1.0	12,432	1.3	19,025	19,171	0.8	19,124	0.5
1966	11,640	11,721	0.7	11,743	0.9	18,000	18,297	1.7	17,858	-0.8
1965	10,800	10,876	0.7	10,948	1.4	16,695	17,071	2.3	16,806	0.7
1964	10,201	10,415	2.1	10,465	2.6	15,788	16,088	1.9	15,924	0.9
1963	9,969	9,980	0.1	9,981	0.1	15,144	15,400	1.7	15,315	1.1
1962	9,500	9,504	-	9,558	0.6	14,900	14,928	0.2	14,950	0.3
1961	9,035	9,120	0.9	9,169	1.5	14,600	14,676	0.5	14,756	1.1
1960	8,800	8,796	-	8,849	0.6	13,536	13,756	1.6	13,983	3.3
1959	8,380	8,393	0.2	8,424	0.5	12,800	13,057	2.0	13,255	3.6
1958	7,800	7,776	-0.3	7,864	0.8	12,000	12,165	1.4	12,206	1.7
Average absolute % difference			0.9		1.2			1.1		1.9
Maximum % diff			-3.1		2.6			2.3		3.6
Number times better			13		-			11		6
Maximum less minimum			5.2		2.5			2.3		7.1

- Rounds to zero.

SOURCE: CURRENT POPULATION SURVEY
U.S. BUREAU OF THE CENSUS

Table 2.--COMPARISON OF DATA ON SELECTED INCOME QUANTILES FOR UNRELATED INDIVIDUALS BY TOTAL MONEY INCOME IN 1958 TO 1974, BY TYPE OF ESTIMATION METHOD, FOR THE UNITED STATES

Year	Ungrouped Data (1)	Pareto-Linear (2)	Percent Difference $\frac{(2)-(1)}{(1)}\times 100$ (3)	Karup-King Log-Log Scale (4)	Percent Difference $\frac{(4)-(1)}{(1)}\times 100$ (5)	Ungrouped Data (6)	Pareto-Linear (7)	Percent Difference $\frac{(7)-(6)}{(6)}\times 100$ (8)	Karup-King Log-Log Scale (9)	Percent Difference $\frac{(9)-(6)}{(6)}\times 100$ (10)
	TWENTIETH PERCENTILE					SIXTIETH PERCENTILE				
1974	$2,095	$2,120	1.2	$2,124	1.4	$5,636	$5,749	2.0	$5,737	1.8
1973	1,872	1,883	0.6	1,891	1.0	5,160	5,242	1.6	5,251	1.8
1972	1,596	1,604	0.5	1,608	0.8	4,660	4,698	0.8	4,680	0.4
1971	1,461	1,472	0.8	1,473	0.8	4,332	4,422	2.1	4,401	1.6
1970	1,368	1,361	-0.5	1,366	-0.1	4,100	4,191	2.2	4,174	1.8
1969	1,235	1,247	1.0	1,255	1.6	3,895	3,906	0.3	3,900	0.1
1968	1,180	1,185	0.4	1,193	1.1	3,600	3,667	1.9	3,651	1.4
1967	1,000	1,015	1.5	1,016	1.6	3,128	3,249	3.9	3,240	3.6
1966	998	998	-	998	-	3,000	3,095	3.2	3,108	3.6
1965	900	870	-3.3	856	-4.9	2,995	2,995	-	2,995	-
1964	839	769	-8.3	748	-10.8	2,654	2,740	3.2	2,727	2.8
1963	792	709	-10.5	685	-13.5	2,400	2,421	0.9	2,407	0.3
1962	775	749	-3.4	756	-2.5	2,340	2,367	1.2	2,350	0.4
1961	695	664	-4.5	664	-4.5	2,340	2,379	1.7	2,364	1.0
1960	650	644	-0.9	645	-0.8	2,400	2,408	0.3	2,399	-
1959	600	568	-5.3	568	-5.3	2,080	2,148	3.3	2,136	2.7
1958	550	559	1.6	559	1.6	2,040	2,128	4.3	2,121	4.0
Average absolute % difference			2.6		3.1			1.9		1.6
Maximum % diff			-10.5		-13.5			4.3		4.0
Number times better			9		3			2		14
Maximum less minimum			12.1		15.1			4.3		4.0
Year	EIGHTIETH PERCENTILE					NINETY-FIFTH PERCENTILE				
1974	$9,296	$9,384	0.9	$9,395	1.1	$15,658	$15,849	1.2	$15,815	1.0
1973	8,802	8,853	0.6	8,860	0.7	15,000	15,216	1.4	15,192	1.3
1972	8,000	8,045	0.6	8,050	0.6	13,500	13,710	1.6	13,775	2.0
1971	7,500	7,528	0.4	7,555	0.7	12,900	12,918	0.1	12,953	0.4
1970	7,200	7,254	0.8	7,281	1.1	12,270	12,435	1.3	12,428	1.3
1969	6,635	6,717	1.2	6,743	1.6	11,800	11,909	0.9	11,917	1.0
1968	6,250	6,375	2.0	6,405	2.5	10,770	10,937	1.6	10,990	2.0
1967	5,593	5,727	2.4	5,756	2.9	9,840	9,925	0.9	9,928	0.9
1966	5,200	5,320	2.3	5,350	2.9	9,200	9,352	1.7	9,372	1.9
1965	5,101	5,260	3.1	5,297	3.8	8,727	8,842	1.3	8,847	1.4
1964	4,996	4,997	-	4,998	-	8,160	8,338	2.2	8,343	2.2
1963	4,675	4,710	0.7	4,748	1.6	8,000	8,074	0.9	8,076	1.0
1962	4,560	4,603	0.9	4,615	1.2	7,800	7,824	0.3	7,834	0.4
1961	4,300	4,373	1.7	4,382	1.9	7,200	7,315	1.6	7,280	1.1
1960	4,181	4,261	1.9	4,277	2.3	6,611	6,753	2.1	6,761	2.3
1959	3,891	3,929	1.0	3,936	1.2	6,492	6,597	1.6	6,605	1.7
1958	3,800	3,836	0.9	3,848	1.3	6,300	6,481	2.9	6,495	3.1
Average absolute % difference			1.3		1.6			1.4		1.5
Maximum % diff			3.1		3.8			2.9		3.1
Number times better			15		-			11		3
Maximum less minimum			3.1		3.8			2.8		2.2

- Rounds to zero.

SOURCE: CURRENT POPULATION SURVEY
U.S. Bureau of the Census

Table 3.--GINI INDEXES AND SELECTED PERCENTAGE SHARES OF AGGREGATE MONEY INCOME IN 1967 TO 1972, FOR ALL FAMILIES, BY TYPE OF ESTIMATION METHOD, FOR THE UNITED STATES

Year	Ungrouped Data (1)	Pareto-Linear (2)	Difference (2)-(1) (3)	Hermite[1] (4)	Difference (4)-(1) (5)	Ungrouped Data (6)	Pareto-Linear (7)	Difference (7)-(6) (8)	Hermite[1] (9)	Difference (9)-(6) (10)
	Gini Index					Lowest 20 Percent				
1972..	.360	.357[2]	-.003	.359	-.001	5.4	5.5	0.1	5.6	0.2
1971..	.356	.355[2]	-.001	.352	-.004	5.5	5.5	-	5.7	0.2
1970..	.354	.353[2]	-.001	.349	-.005	5.4	5.5	0.1	5.7	0.3
1969..	.349	.347[2]	-.002	.345	-.004	5.6	5.6	-	5.8	0.2
1968..	.348	.344[2]	-.004	.342	-.006	5.6	5.7	0.1	5.9	0.3
1967..	.348	.347[2]	-.001	.344	-.004	5.5	5.6	0.1	5.8	0.3
Year	60 TO 80 PERCENT					TOP 5 PERCENT				
1972..	23.9	23.8	-0.1	23.7	-0.2	15.9	15.9	-	16.2	0.3
1971..	23.8	23.8	-	23.7	-0.1	15.7	15.9	0.2	15.6	-0.1
1970..	23.8	23.8	-	23.7	-0.1	15.6	15.8	0.2	15.4	-0.2
1969..	23.7	23.7	-	23.7	-	15.6	15.6	-	15.4	-0.2
1968..	23.7	23.8	0.1	23.7	-	15.6	15.3	-0.3	15.4	-0.2
1967..	23.9	23.8	-0.1	23.8	-0.1	15.2	15.3	0.1	15.4	0.2

- Rounds to zero.

1 Gastwirth and Glauberman, "On the Interpolation of the Lorenz Curve and Gini Index", Unpublished Paper.

2 Gini Index calculated using Simpson's rule for approximate integration after splitting the Lorenz Curve into 100 equal intervals.

SOURCE: CURRENT POPULATION SURVEY
U.S. BUREAU OF THE CENSUS

Table 4.--GINI INDEX AND PERCENTAGE SHARE OF AGGREGATE MONEY INCOME IN 1958 TO 1974 RECEIVED BY THE TOP 5 PERCENT OF FAMILIES AND UNRELATED INDIVIDUALS, FOR THE UNITED STATES

Year	FAMILIES						UNRELATED INDIVIDUALS					
	Gini Index			Top 5 Percent			Gini Index			Top 5 Percent		
	Ungrouped Data	Pareto-Linear[1]	Difference	Ungrouped Data	Pareto-Linear	Difference	Ungrouped Data	Pareto-Linear[1]	Difference	Ungrouped Data	Pareto-Linear	Difference
1974..	.356	.352	-.004	15.3	15.4	0.1	.448	.446	-.002	19.3	19.4	0.1
1973..	.357	.355	-.002	15.5	15.8	0.3	.460	.463	.003	20.0	20.9	0.9
1972..	.360	.357	-.003	15.9	15.9	-	.478	.474	-.004	21.4	21.3	-0.1
1971..	.356	.355	-.001	15.7	15.9	0.2	.473	.471	-.002	20.5	20.5	-
1970..	.354	.353	-.001	15.6	15.8	0.2	.478	.478	-	20.8	20.9	0.1
1969..	.349	.347	-.002	15.6	15.6	-	.481	.478	-.003	20.7	20.6	-0.1
1968..	.348	.344	-.004	15.6	15.3	-0.3	.480	.478	-.002	20.8	20.4	-0.4
1967..	.348	.347	-.001	15.2	15.3	0.1	.490	.491	.001	21.1	21.2	0.1
1966..	.349	.348	-.001	15.6	15.6	-	.484	.488	.004	21.2	21.4	0.2
1965..	.356	.356	-	15.5	15.7	0.2	.486	.487	.001	20.0	20.1	0.1
1964..	.361	.356	-.005	15.9	15.4	-0.5	.512	.508	-.004	22.9	22.3	-0.6
1963..	.362	.359	-.003	15.8	15.6	-0.2	.500	.504	.004	20.1	21.0	0.9
1962..	.362	.362	-	15.7	15.9	0.2	.502	.497	-.005	20.8	21.0	0.2
1961..	.374	.373	-.001	16.6	16.8	0.2	.510	.508[2]	-.002[2]	21.6	22.4[2]	0.8[2]
1960..	.364	.366	.002	15.9	16.6	0.7	.506	.490[2]	-.016[2]	20.2	19.9[2]	-0.3[2]
1959..	.361	.360	-.001	15.9	16.1	0.2	.522	.524	.002	22.1	24.4	2.3
1958..	.354	.354	-	15.4	15.6	0.2	.519	.505	-.014	21.6	21.3	-0.3

- Rounds to zero.

1 Gini Index calculated using Simpson's rule for approximate integration after splitting the Lorenz Curve into 100 equal intervals.

2 Pareto invalid in top interval. Assumed Pareto Alpha = 2.85.
Mean for $25,000 and over = $37,500, mean for $15,000 to $25,000 = $20,000.

SOURCE: CURRENT POPULATION SURVEY
U.S. BUREAU OF THE CENSUS

A SIMPLIFIED URBAN HOUSING INVENTORY MODEL - WITH PRACTICAL APPLICATIONS

Ko Ching Shih, U.S. Department of Housing & Urban Development

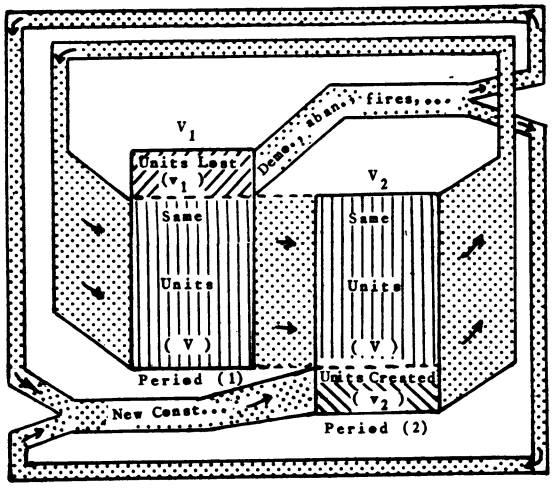
I. Introduction

Since 1950, the Bureau of the Census has established a standard procedure for measuring the changes of housing inventory components for any given place in the United States (1). In the early 1960's, the economic staff of the Federal Housing Administration utilized the Census Bureau's procedure extensively as part of the FHA's official housing market analysis techniques (2), and applied them to many housing market areas throughout the nation.

II. A Macro-model

Housing inventory can be modeled as in Figure 1:

Figure 1. A Macro-model of Housing Inventory - General View



In this model, housing units are distributed into three basic components:

- V - those units that are common to both time periods 1 and 2
- v_1 - those units that were lost or removed between the last and the current inventory counts
- v_2 - those units that were added or created between the last and the current inventory counts

Aggregately, the total housing inventory of the current period is

$$V_2 = (V_1 - v_1) + v_2 \quad (1)$$

or equivalently,

$$V_2 = V_1 + (v_2 - v_1) \quad (2)$$

$(v_2 - v_1)$ is the net inventory change between the two time periods, and the rate of net inventory change is

$$g = \frac{(v_2 - v_1)}{(v_2 + v_1)} \quad (3)$$

The value of g ranges from $-1 \leq g \leq +1$.

Between January 1, 1970 and December 30, 1976, Chicago lost about 21,900 housing units per year and built only 5,500 new units annually (3,4). Thus, at the end of 1976, Chicago had a g value of -0.60. In the same seven-year period, Schaumburg, a new community in suburban Cook County, Illinois, issued about 1,500 building permits annually and lost only about 50 units per year (5), so Schaumburg had a g value of +0.93.

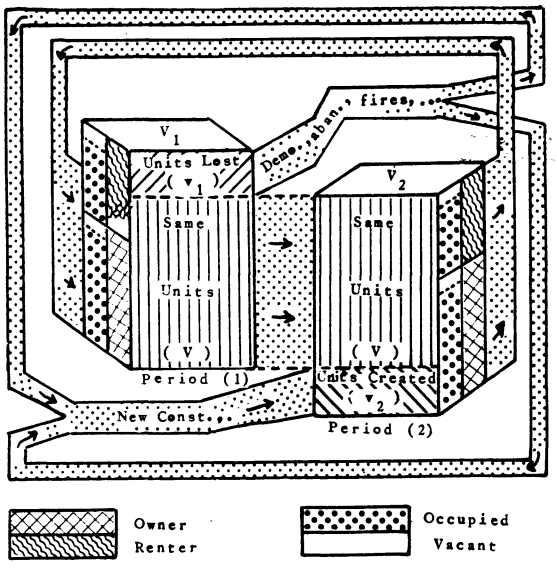
For the estimation of V_2 for a rapidly growing place, the critical stratum is v_2; for a declining central city, the critical stratum is v_1; the critical estimator in both cases is g.

For most urban place in the United States, the value of g ranges from -0.25 to +0.25. Thus, the study of the characteristics of V, which constitutes the major components of both V_1 and V_2, must be carried out in order to yield an unbiased estimate of V_2.

III. A Multidimensional View

The macro-model of the housing inventory is multidimensional. Figure 2 shows the model segmented in terms of tenure and occupancy status.

Figure 2. A Macro-model of Housing Inventory - Tenure and Occupancy Status



The above model can be described by four equations:

$$V = V_o + V_c \tag{4}$$

$$v_2 = v_{2,o} + v_{2,c} \tag{5}$$

$$V_2 = [(V_{ow} + V_{or}) + (V_{cw} + V_{cr})] + [(v_{2,ow} + v_{2,or}) + (v_{2,cw} + v_{2,cr})] \tag{6}$$

$$V_2 = [(V_{ow} + v_{2,ow}) + (V_{cw} + v_{2,cw})] + [(V_{or} + v_{2,or}) + (V_{cr} + v_{2,cr})] \tag{7}$$

where

V_o ≡ occupied units
V_{ow} ≡ owner occupied units
V_{or} ≡ renter occupied units
V_c ≡ vacant units
V_{cw} ≡ vacant units available for sale
V_{cr} ≡ vacant units available for rent
$v_{2,o}$ ≡ new units occupied
$v_{2,c}$ ≡ new units vacant
$v_{2,ow}$ ≡ new sales units occupied by owners
$v_{2,or}$ ≡ new rental units occupied by renters
$v_{2,cw}$ ≡ new sales units available for sale
$v_{2,cr}$ ≡ new rental units available for rent

$(V_{ow} + v_{2,ow})$ is the approximate number of current homeowners, and $(V_{or} + v_{2,or})$ is the estimated current number of renters. The current number of residential households, H_2, is

$$H_2 = (V_{ow} + v_{2,ow}) + (V_{or} + v_{2,or}) \tag{8}$$

For a given place, the average size of a household could be estimated by a small stratified survey as defined by equation (8), or by the least squares method if time series data is available.

The current aggregate population could also be easily estimated by

$$P_2 = \alpha H_2 \tag{9}$$

where α is the estimated size of a residential household. The rate of new household formation is

$$h = \frac{v_{2,ow} + v_{2,or}}{(V_{ow} + v_{2,ow}) + (V_{or} + v_{2,or})} \tag{10}$$

h is a critical estimator for projecting the number of residential households and the total residential population, particularly for a rapidly growing place.

$[(V_{ow} + v_{2,ow}) + (V_{cw} + v_{2,cw})]$ is the homeowner inventory, and the homeowner vacancy rate is

$$c_w = \frac{(V_{cw} + v_{2,cw})}{(V_{ow} + v_{2,ow}) + (v_{cw} + v_{2,cw})} \tag{11}$$

Accordingly, three additional equations may be deduced:

$$v_{2,u} = \frac{v_{2,cw}}{(v_{2,ow} + v_{2,cw})} \tag{12}$$

$$c_r = \frac{(V_{cr} + v_{2,cr})}{(V_{or} + v_{2,or}) + (V_{cr} + v_{2,cr})} \tag{13}$$

$$y = 1 - c_r \tag{14}$$

where

$v_{2,u}$ ≡ unsold new home inventory ratio
c_r ≡ rental vacancy rate
y ≡ rental occupancy factor

In many larger urban areas in the United States, most of the new single-family sales units are concentrated in new subdivisions. The FHA and local homebuilder organizations survey these unsold new units annually. Thus for the estimation of vacant sales housing, the critical strata are new subdivisions, and the critical estimator is the unsold inventory ratio.

In large urbanized areas, many of the rental units are concentrated in garden type projects or high-rise complexes; all of these larger rental projects are managed by specialized firms who usually compute monthly occupancy factors. Thus for the estimation of the rental vacancy rate, the critical strata are those neighborhoods or blocks with high concentrations of multifamily rental structures, J_m, and the critical estimator is y.

A series of equations for each dimension of the macro-model could be written. Following are a series for the assessment of housing quality (6) :

$$a = \frac{v_{1,a}}{v_1} \tag{15}$$

$$q = f(t, m) \tag{16}$$

$$q_s = 1 - \sum_j \Delta q_j \tag{17}$$

$$r = \frac{\partial f}{\partial t} \tag{18}$$

where

a ≡ abandonment ratio

$v_{1,a}$ ≡ aggregate units abandoned in previous period

q ≡ quality coefficient of a housing structure as a function of time t and maintenance level m

q_s ≡ quality coefficient of housing inventory at substandard point s

r ≡ rate of substandardization

IV. Critical Strata and Estimators

Statistically, each dimension of the macro-model consists of one or more critical strata and corresponding critical estimators. For the purpose of generating the most reliable estimates of various urban variables, these critical strata must be identified and controlled during the development of a sampling frame, the establishment of a data system, the execution of multistage stratified probability sampling, and during the control of sampling and non-sampling errors. Critical strata and estimators are summarized in Table 1.

Table 1. Summary of Selected Critical Strata and Estimators

Variable	Critical Strata	Critical Estimator
V_2	v_2, v_1	g
c_w	J_n	$v_{2,u}$
c_r	J_m	y
H_2	v_2, v_1	g
P_2	v_2, v_1	α
a	J_{v_1}	g
q_s	J_{v_1}	r

where

J_n ≡ new subdivisions

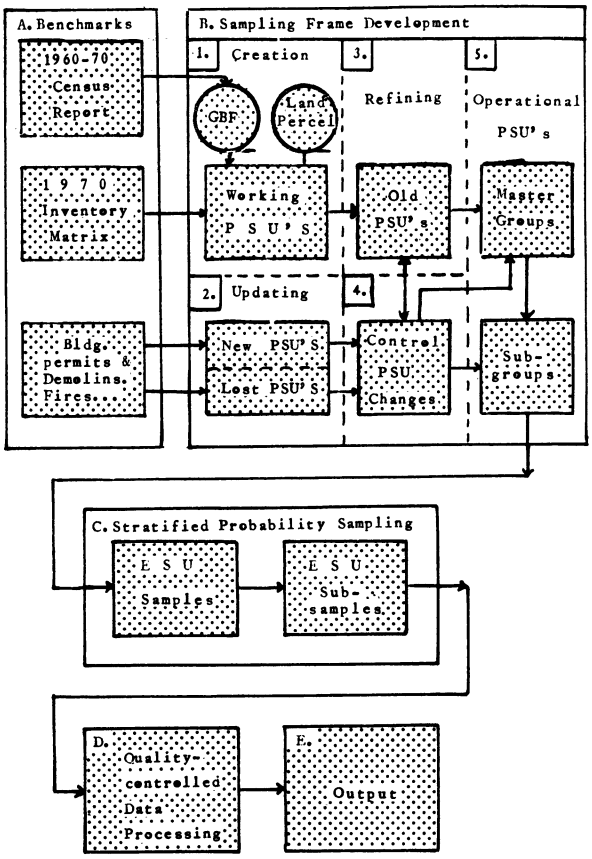
J_m ≡ neighborhoods with concentrations of large multifamily structures

J_{v_1} ≡ neighborhoods with concentrations of inventory loss

V. A Data System

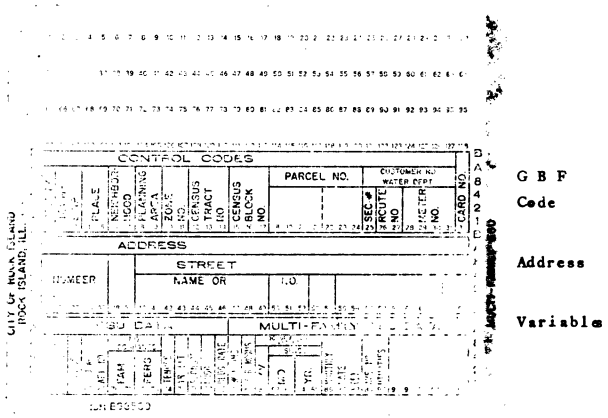
From the statistician's point of view, an efficient data system must be capable of stratifying PSU's into desirable groups and subgroups which can be operated either independently or jointly in order to maximize sampling efficiency. A condensed version of a simplified housing inventory data system is shown in figure 3 (7).

Figure 3. A Simplified Housing Inventory System



The data system is adaptable to any level of automation. Figure 4 shows an example of a PSU unit record for the Rock Island, Illinois system.

Figure 4. PSU Unit Record, Rock Island, Illinois System



The system consists of five components:

A. Benchmarks -- benchmarks insure that the final output is statistically comparable with the latest available Census inventory matrix. Many urban places with a population of 5000 or more are in the Census Bureau's samples of permit-issuing and demolition surveys. Thus with some data collection on fire and other losses, a time series of g values could be estimated. Most communities in the U.S. have a building department that issues permits for new construction, demolition, and conversions. This is the main source for v_1 and v_2 data. A standard unit record input device such as the one shown in Figure 4 (8) is the updating subsystem for the development of a comprehensive sampling frame.

B. Development of a Sampling Frame -- the sampling frame is the key component of the system. The objective in the development of the sampling frame was to maintain operational flexibility and high reliability. Based on a geographical base file (GBF) or an existing land-use parcel file, or any directory of buildings, a working PSU for existing housing structures could be created. Using a predetermined sampling ratio and procedure, working PSU's were selected on a rotating basis over a fixed time period. They were then stratified into subgroups for refinement and analysis. Refined PSU's were then regrouped into a predetermined number of operational PSU's and sub-PSU's.

C. Multistage Stratified Probability Sampling-- ESU's were randomly selected in several stages from either refined PSU's or sub-PSU's. In practice, the size and other features of ESU's are determined by the requirements of the output matrix and their prescribed confidence levels.

D. Quality Control -- critical estimators play a significant role in this component.

E. Output -- a variety of output matrices are available, including the computed sampling error tables.

The primary features of the data system are

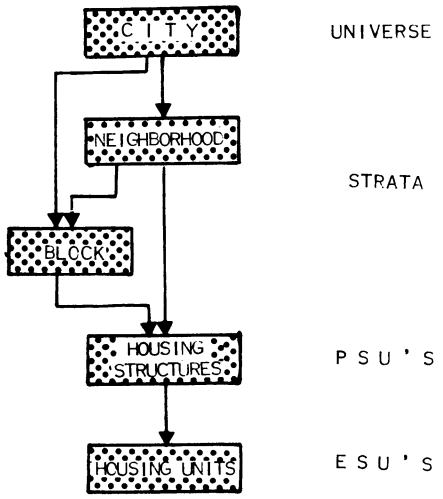
- staged development of a series of desirable PSU's
- refined treatment of the developed PSU's

- dynamic maintenance of a series of independent sub-PSU's
- Flexibility of multistage stratified probability samplings and control
- high reliability at a relatively low cost
- a multitude of applications because of the interchangability of ESU's and households

VI. Procedures

A. Sampling Process -- the successive elements involved in the sampling process are shown in Figure 5.

Figure 5. Elements of the Sampling Process



In certain special cases, city blocks are used as strata even though only a small sample is required. For example in Chicago, Illinois, the absorption rate of high-rise condominums is estimated using city blocks as strata because most units are concentrated along the lake shore.

B. Sampling Plan -- In many urban areas, the distribution of PSU's in terms of the size of structure is quite significant. Therefore disproportionate cut-off sampling is the method of choice.

C. Control of Non-sampling Errors -- because of the extensive refinement and stratification of PSU's, non-response recall, survey control, and quality control editing of questionaire returns could be efficiently executed. Non-sampling errors are therefore controlled, and costs may be reduced.

E. Quality of Output -- since most of the critical estimators are known, the quality of output will be comparatively high and statistically acceptable.

VII. Potential Applications

In urban areas, housing constitutes the most significant sector of land use. A comprehensive housing inventory model is therefore the major component of a total urban planning model. It generally covers most of the variables involved in the measurement of urban planning and programming adequacies, cost-benefit analyses, allocation of limited resources, projection of transportation and community facility requirements, and the development and implementation of urban socioeconomic models. The Rock Island, Illinois, Total Housing Inventory System (8) was developed with these long term objectives in mind.

The Rock Island data base covers every piece of land in the city, including vacant parcels. Thus the system can generate much desirable time series data on a broad spectrum of urban variables in addition to serving the requirements of a housing inventory system. It is considered to be a comprehensive version of an urban housing inventory model.

Alternatively, a simplified housing inventory model could be developed based on almost any acceptable data base as shown in Figure 3, and be maintained at comparably less cost with some advantageous features.

The federal, state, and local governments have collectively spent a large sum on a variety of urban programs. Many of these programs were adopted with little or no testing or empirical data, mainly because of the lack of a current dynamic sampling frame. If a series of simplified housing inventory models were developed and maintained at strategic locations, many of the hypotheses of urban programs could be tested on short notice. A significant contribution to the decision-making process could result.

In addition, if a network of such housing inventory models is maintained, not only will the communities involved benefit in their daily operations, but the system could be utilized as an urban research laboratory. Urban planners, researchers, and governmental and non-governmental agencies could utilize the lab for

- testing of hypotheses of proposed new urban programs
- testing of significance of differences between competing program proposals
- evaluation of the performance of existing programs
- simulation or testing of developed urban models
- testing new survey questionnaires and procedures

VIII. References

1. U.S. Bureau of the Census, Census of Housing, 1970, Volume IV, Components of Inventory Change, Washington (1973)

2. U.S. Department of Housing and Urban Development, FHA Techniques of Housing Market Analysis, Washington (1970)

3. City of Chicago, Department of Development and Planning, Annual Housing Report, Chicago (1975, 1976)

4. City of Chicago, Housing Assistance Plan, Chicago (1977)

5. U.S. Bureau of the Census, Construction Report - Housing Authorized by Building Permits, Washington (1960 - 1977, monthly)

6. Ko Ching Shih, "Measuring the Quality of Housing, " Proceedings of Social Statistics Section, American Statistical Association, 358-363 (1971)

7. A detailed description of the data system is too large for publication. Limited copies are available upon written request to the author

8. City of Rock Island, Illinois, Total Housing Inventory System, Rock Island, Illinois (1971)

Selected Characteristics of Persons Who Reported a Work Disability in the 1970 Census

John M. McNeil and Douglas K. Sater, Bureau of the Census

Introduction

The economic, demographic and social characteristics of persons who perceive themselves to be work disabled differ considerably from the characteristics of those who do not. This paper uses data from the 1970 census to identify and measure some of these differences. Also, because of the interest in developing estimates of disability rates for local areas, state level data are used to examine the association between disability rates and such variables as income, age, education levels, and industrial and occupational structure.

Data Source

The data presented in this paper on the characteristics of disabled and nondisabled persons are based on a special tabulation of the 1-in-100 and the 1-in-1,000 public use samples of the 1970 census. The data on state disability rates and other state characteristics are taken from published sources. The 1970 census work disability questions asked persons 14 to 64 years of age whether they were limited in the kind or amount of work they could do, whether they could work at any job at all, and for how long had they been limited in their ability to work. Persons were classified as "completely disabled" if they were unable to work at any job at all; "partially disabled" if they were able to work at a job, but were limited in the kind or amount of work they could do; and "not disabled" if they were not limited in the kind or amount of work they could do.

Sampling Variability

The work disability questions were asked in a 5 percent sample of households. In addition, the public use files are a representative subsample of these households. Thus, the data from this source are subject to errors due to sampling variability. The standard errors for numbers and percents are not shown, but they have been computed and comparisons will be made in the text only if the differences exceed a level that could be attributable to sampling error.

Nonsampling Error

An individual's response to a survey question on work disability status is necessarily subjective. The phrase "limited in the kind or amount of work he can do" is open to a wide range of interpretations and even the concept of a complete work disability is not unambiguous. An individual's response to a work disability question may be determined by factors other than the actual physical or mental condition of the person. That is, persons with similar medical problems may differ considerably in their desire to work, in their education and training and in their ability or opportunity to adapt to particular work situations. Thus, care should be exercised in the interpretation of differences between the disabled and nondisabled populations.

Sex

There was a substantial difference between males and females in the percent reporting a work disability. About 8.1 percent of all males aged 18 to 64 have a partial work disability compared with 4.9 percent of all females. However, only 3.8 percent of the males report a complete work disability. This compares with 5.1 percent of the females. It seems reasonable to suppose that most of the difference is due to factors that affect labor force attachment. That is, while males tend to be more aware of their work limitations, they are historically more strongly attached to the labor force.

Race and Poverty

Black persons and persons in poverty had a higher incidence of complete work disability than White persons and persons above the poverty level. Among Blacks the figure was 7.6 percent and among Blacks in poverty the figure was 15.0 percent. The comparable figures for Whites and Whites in poverty were 4.2 percent and 14.0 percent.

Schooling

Persons who report a partial or complete work disability have, on the average, completed fewer years of schooling than nondisabled persons. About 62.9 percent of those persons with no work disability completed 12 or more years of schooling while 48.5 percent of those with a partial work disability and only 28.1 percent of those with a complete work disability completed 12 or more years of schooling.

Marital Status

Persons with a work disability were more likely to be separated, widowed or divorced than were persons with no work disability. The percent of nondisabled males who were in one of the three categories was about 5.3 percent compared to about 14.7 percent for males with a complete work disability. Among nondisabled females, 12.3 percent were separated, widowed or divorced. Among females with a complete work disability, the figure was 27.3 percent.

Persons with a complete work disability were less likely than nondisabled persons to live with an employed spouse. About 30.6 percent of all nondisabled males and about 23.6 percent of all completely disabled males lived with an employed spouse. The comparable figures for females were 64.8 percent and 41.7 percent.

Personal and Family Income

Persons with a complete work disability had substantially lower personal and family incomes than persons with no work disability. Completely disabled males had about 33.4 percent of the mean personal income and about 51.4 percent of the mean family income of nondisabled males. Mean personal income was $8,481 for nondisabled males and $2,832

for completely disabled males. The mean income of other family members was approximately $3,500 for both disabled and nondisabled males. Thus, completely disabled males contributed, on the average, about 45.8 percent of their family income, while nondisabled males contributed about 70.4 percent. For females, the differences in income associated with work disability status were slightly smaller. That is, completely disabled females had about 38.8 percent of the mean personal income and about 62.9 percent of the mean family income of nondisabled females. Nondisabled females had a mean personal income of $2,385 and a mean family income of $11,208. The comparable figures for completely disabled females were $925 and $7,045. Thus, completely disabled females contributed, on the average, about 13.1 percent of their family income, while nondisabled females contributed about 21.3 percent.

Income Sources

About 29.5 percent of the males who reported a complete work disability in the 1970 census reported that they had received some earnings in 1969. This compares with about 95.5 percent of the nondisabled males. About 39.6 percent of work disabled males reported the receipt of income from Social Security or Railroad Retirement, 15.3 percent reported receiving public assistance and 32.2 percent reported income from other sources. The comparable figures for the nondisabled males were 1.5 percent, 0.7 percent, and 18.5 percent. Completely disabled females were about as likely as completely disabled males to have received public assistance but were much less likely to have received earnings, income from Social Security and Railroad Retirement or income from other sources.

Another measure of interest is the percent of income accounted for by a particular source. In general, disabled persons who received Social Security or Railroad Retirement, public assistance, or income from other sources tended to rely more on that income than nondisabled persons. Income from public assistance accounted for 33.5 percent of the total income of those nondisabled male family heads who received such income. The comparable figure for males with a complete work disability was 66.0 percent. For females, the comparable figures were 73.7 percent for family heads with no disability and 83.6 percent for completely disabled family heads.

Earnings of Workers With a Work Disability

In the process of developing a model that would examine male-female earnings differentials[2] in 1970, the earnings of about 51,000 persons who worked in 1969 were regressed on: Age, education, income of other family members, sex, age at first marriage, class of worker, activity five years ago, hours and weeks worked, and work disability status. Each variable was either recoded into a suitable variable or a complete set of dichotomous variables. The disability variable was given a code of "1" for a partial or complete work disability, and "0" for not disabled.

All variables were in the final model and were significant ($\alpha = .05$) except for a few of the age by education dummy variables. The $\bar{R}^2$ of the final model was .452. The coefficient of the work disability variable was -823.8 with a "t" statistic of -10.1. That is, over and above the differences explained by other variables, persons who had earnings and who had a work disability had $824 less annual earnings than those with no work disability. Because of the correlation between work disability and education, the work disability coefficient actually understates the relationship between work disability and earnings.

State Variations

There is considerable State by State variation in the proportion of persons with a work disability. The work disability rates tend to be low in the northeastern States and high in the southern States. Alaska and Hawaii have very low work disability rates. The following section reports on a preliminary attempt to identify factors that are associated with these variations.

The 26 variables selected for their possible association with work disability rates and their corresponding simple correlation coefficients are shown in table 4. The 6 variables that have the highest absolute correlation with the percent of persons reporting a complete work disability are the percent of persons receiving Social Security benefits, the relative level of Social Security disability benefits, the percent of families in poverty, the percent of unrelated individuals in poverty, median family income, and median school years completed. It should be noted that there is a significant degree of intercorrelations among the variables. For example, the two schooling variables, the white-collar worker variable, and the percent employed in construction variable are all highly correlated with income. This obviously affects the interpretation of the coefficients because of the proxy representation of other factors. As a technical note, the regression package we used calculates estimates using the rel-variance, rel-covariance matrix. Thus, even with larger intercorrelations than we incurred, the coefficient estimates will still be relatively accurate.

Table 5 shows results from two equations based on a step-wise regression procedure. In the first equation, the dependent variable was defined to be the percent of persons in the state with either a partial or a complete work disability. In the second equation, the dependent variable was defined to be the percent with a complete work disability.

The proportion of the variance explained is not great in either equation. The $\bar{R}^2$ is .76 when the dependent variable is the percent with either a partial or a complete work disability, and .88 when the dependent variable is the percent of persons with a complete work disability. The most significant independent variable in the first equation is the unemployment rate. That is, the higher the unemployment rate, the higher was the reported work disability rate. The other variables that entered were median family income, the

percent of persons in poverty and the percent of employed persons in manufacturing.

In the second equation, the percent of persons in poverty is the most significant independent variable by a wide margin. The other entering variables are the percent receiving public assistance, the percent receiving Social Security, the percent employed in coal mining, the percent of employed persons in agriculture (with a negative sign), and the percent of employed persons in white-collar occupations.

The results of the regression study are not particularly impressive, partly because of the crude way in which the industrial and occupational factors were defined. But, because an equation with a high degree of explanatory power would be useful in making synthetic estimates of the prevalence of disability in various areas, we expect to continue to work in this area.

FOOTNOTES

1/ For a detailed discussion of the sample design, editing, allocation, estimate and sampling variability, see appendix C in:

U.S. Bureau of the Census
Census of Population: 1970
General Social and Economic Characteristics
Final Report PC(1)-C2 through C52

2/ McNeil, Jack and Douglas Sater. "Recent Changes in Female to Male Earnings Ratios" Paper presented at the Population Association Meeting in Seattle in April 1975.

Table 1. -- Distribution of Persons 18 to 64 Years of Age by Work Disability Status, Age, Sex, Race and Poverty Status: 1970

(Numbers in thousands)

Characteristics	Total	Not disabled		Partially disabled		Completely disabled	
		Number	Horizontal percent	Number	Horizontal percent	Number	Horizontal percent
AGE							
Total persons 18 to 64 years of age...	108,305	96,472	89.1	6,950	6.4	4,884	4.5
18 to 44 years of age.....	67,089	62,585	93.3	3,112	4.6	1,391	2.1
45 to 54 years of age.....	22,756	19,589	86.1	1,904	8.4	1,263	5.6
55 to 59 years of age.....	9,875	7,904	80.0	1,008	10.2	963	9.8
60 to 64 years of age.....	8,585	6,393	74.5	926	10.8	1,266	14.7
RACE AND POVERTY STATUS							
Total.....	108,305	96,472	89.1	6,950	6.4	4,884	4.5
Poor.....	10,768	8,292	77.0	951	8.8	1,526	14.2
White.....	96,137	86,029	89.5	6,102	6.3	4,007	4.2
Poor.....	7,704	5,942	77.1	680	8.8	1,081	14.0
Black.....	10,771	9,176	85.2	773	7.2	820	7.6
Poor.....	2,815	2,136	75.9	256	9.1	422	15.0
SEX							
Males.....	51,505	45,351	88.1	4,185	8.1	1,970	3.8
Females.....	56,800	51,121	90.0	2,765	4.9	2,914	5.1

Table 2. -- Distribution of Persons 18 to 64 Years of Age by Work Disability Status, Sex, Marital Status and Education: 1970

(Numbers in thousands)

Characteristics	Total	Not disabled		Partially disabled		Completely disabled	
		Number	Vertical percent	Number	Vertical percent	Number	Vertical percent
SEX AND MARITAL STATUS							
Males.....	51,505	45,351	100.0	4,185	100.0	1,970	100.0
Married, wife present.....	38,424	34,142	75.3	3,089	73.8	1,193	60.6
Wife employed.....	15,696	13,887	30.6	1,344	32.1	465	23.6
Married, wife absent.....	765	668	1.5	61	1.5	36	1.8
Widowed, divorced, separated....	3,009	2,394	5.3	325	7.8	290	14.7
Never married.....	9,307	8,148	18.0	709	16.9	450	22.8
Females.....	56,800	51,121	100.0	2,765	100.0	2,914	100.0
Married, husband present.....	40,149	36,660	71.7	1,753	63.4	1,736	59.6
Husband employed.....	35,778	33,112	64.8	1,450	52.4	1,215	41.7
Married, husband absent.....	1,145	1,038	2.0	57	2.1	50	1.7
Widowed, divorced, separated....	7,727	6,275	12.3	657	23.8	796	27.3
Never married.....	7,779	7,148	14.0	298	10.8	333	11.4
HIGHEST GRADE COMPLETED							
Total persons.....	108,305	96,472	100.0	6,950	100.0	4,884	100.0
Under 8 years completed.....	10,829	8,158	8.5	1,103	15.9	1,569	32.1
8 to 11.....	32,073	27,653	28.7	2,477	35.6	1,942	39.8
12 or more.....	65,403	60,660	62.9	3,370	48.5	1,373	28.1
16 or more.....	11,712	11,068	11.5	513	7.4	131	2.7

Table 3. -- Persons Receiving Income From Various Sources by Work Disability Status, Sex, and Family Relationship: 1970

Characteristics	Males				Females			
	Total	Not disabled	Partially disabled	Completely disabled	Total	Not disabled	Partially disabled	Completely disabled
MEAN PERSONS INCOME								
All persons	$ 8,147	$ 8,481	$ 7,024	$2,832	$ 2,298	$ 2,385	$ 2,128	$ 925
Married, spouse present	9,413	9,742	8,052	3,549	1,852	1,919	1,691	617
MEAN FAMILY INCOME								
All persons	$11,696	$12,045	$10,507	$6,186	$10,896	$11,208	$ 9,187	$7,045
Married, spouse present	12,051	12,348	10,909	6,545	11,922	12,137	10,843	8,479
NUMBER RECEIVING INCOME BY RELATIONSHIP AND SOURCE OF INCOME (In thousands)								
All persons	51,505	45,351	4,185	1,970	56,800	51,121	2,765	2,914
Earnings	47,770	43,331	3,859	580	31,481	29,512	1,545	424
Social Security or Railroad Retirement	1,661	670	210	781	2,618	1,794	229	595
Public assistance	752	331	119	302	1,669	1,079	159	431
Other sources	10,088	8,378	1,077	635	5,206	4,497	348	361
Family heads	39,103	34,683	3,176	1,245	4,603	3,911	357	335
Earnings	37,461	34,036	3,012	413	3,168	2,878	233	57
Social Security or Railroad Retirement	1,136	418	146	572	701	534	65	102
Public assistance	497	246	83	168	843	617	85	141
Other sources	8,579	7,188	914	477	1,107	949	84	74
Other family members	7,896	6,858	601	438	47,255	43,012	2,055	2,188
Earnings	6,364	5,778	494	92	24,489	23,123	1,060	306
Social Security or Railroad Retirement	295	154	32	108	1,403	945	109	349
Public assistance	126	42	16	68	593	378	45	170
Other sources	637	500	66	72	2,903	2,573	161	170
Unrelated individuals	4,505	3,811	408	286	4,941	4,198	353	391
Earnings	3,944	3,517	353	75	3,824	3,511	252	61
Social Security or Railroad Retirement	231	98	32	101	514	315	55	144
Public assistance	130	43	20	66	234	84	29	120
Other sources	872	690	97	86	1,144	975	103	117
MEAN INCOME RECEIVED BY SOURCE OF INCOME AND RELATIONSHIP								
All persons:								
Earnings	$ 8,372	$ 8,550	$ 6,981	$4,364	$ 3,678	$ 3,728	$ 3,116	$2,253
Social Security or Railroad Retirement	1,241	1,068	1,200	1,401	1,034	1,065	968	967
Public assistance	1,023	902	1,086	1,131	1,356	1,444	1,277	1,164
Other sources	1,675	1,588	1,892	2,442	1,757	1,762	1,716	1,741
Family heads:								
Earnings	9,378	9,574	7,778	4,935	4,506	4,641	3,392	2,237
Social Security or Railroad Retirement	1,357	1,175	1,290	1,508	1,455	1,529	1,259	1,195
Public assistance	1,096	952	1,182	1,263	1,673	1,727	1,513	1,532
Other sources	1,713	1,625	1,939	2,608	2,039	2,079	1,755	1,850
Other family members:								
Earnings	3,643	3,696	3,257	2,428	3,401	3,438	2,944	2,218
Social Security or Railroad Retirement	917	846	905	1,031	842	830	795	891
Public assistance	802	737	711	863	1,022	1,066	1,011	929
Other sources	1,090	975	1,221	1,750	1,599	1,609	1,535	1,491
Unrelated individuals:								
Earnings	6,449	6,615	5,394	3,594	4,762	4,887	3,584	2,441
Social Security or Railroad Retirement	1,079	963	1,082	1,192	983	983	968	990
Public assistance	953	780	985	1,070	1,052	1,064	1,000	1,064
Other sources	1,724	1,650	1,901	2,104	1,885	1,856	1,968	2,034
MEAN PERCENT OF TOTAL INCOME FOR PERSONS RECEIVING INCOME FROM EACH SOURCE								
Family heads:								
Earnings	97.0	97.5	93.6	84.3	86.5	87.2	81.8	70.5
Social Security or Railroad Retirement	53.3	36.6	45.2	67.5	55.1	51.6	56.3	73.0
Public assistance	46.2	33.5	43.4	66.0	75.0	73.7	70.1	83.6
Other sources	17.1	13.7	22.5	57.8	39.8	37.6	45.2	61.8
Other family members:								
Earnings	97.7	98.0	95.4	91.6	97.9	98.0	95.7	92.6
Social Security or Railroad Retirement	66.0	56.8	59.9	81.5	75.2	71.1	72.7	87.3
Public assistance	73.6	56.6	67.3	85.6	77.7	74.0	72.0	87.4
Other sources	33.8	27.8	37.9	71.8	52.2	50.5	57.1	73.7
Unrelated individuals:								
Earnings	95.6	96.1	92.2	84.8	93.8	94.4	88.1	76.7
Social Security or Railroad Retirement	64.5	60.6	56.4	70.8	59.7	56.0	54.6	69.8
Public assistance	65.7	53.9	57.9	76.8	72.8	64.3	68.5	80.3
Other sources	31.6	26.5	34.9	68.1	40.0	35.7	48.7	67.6

Table 4. -- Weighted Intercorrelation Coefficients Between the Percent of Persons Reporting a Work Disability and Variables Selected for Their Possible Association With Work Disability Rates

	PCTDIS	PCTUNA	MEDSCH	PCTHS4	PCTOLD	PCTPOV	UIPOV	PCTUNP	PCTURB	MEDFIN	PCTPA	PCTSS	PCTWCW	PCTAGR
MEDSCH...	-.58	-.75	1.00											
PCTHS4...	-.41	-.63	.91	1.00										
PCTOLD...	-.10	-.04	.17	.02	1.00									
PCTPOV...	.77	.85	-.85	-.73	-.20	1.00								
UIPOV....	.67	.72	-.82	-.77	.04	.85	1.00							
PCTUNP...	.26	.08	.30	.52	-.12	-.13	-.27	1.00						
PCTURB...	-.48	-.48	.69	.62	.04	-.62	-.82	.23	1.00					
MEDFIN...	-.74	-.75	.74	.65	.02	-.90	-.89	.20	.72	1.00				
PCTPA....	.52	.61	-.30	-.12	-.10	.50	.16	.42	.12	-.28	1.00			
PCTSS....	.23	.20	.01	-.05	.80	.04	.21	-.08	-.13	-.27	-.08	1.00		
PCTWCW...	-.49	-.47	.68	.66	-.02	-.55	-.82	.25	.86	.65	.17	-.17	1.00	
PCTAGR...	.33	.12	-.19	-.05	.04	.38	.44	-.02	-.54	-.53	-.08	.25	-.43	1.00
PCTMIN...	.35	.40	-.32	-.26	-.06	.45	.49	.02	-.31	-.46	.13	.00	-.23	.17
PCTCON...	.47	.48	-.50	-.42	-.32	.69	.56	-.20	-.40	-.69	.15	.03	-.29	.30
PCTMFG...	-.18	-.08	-.17	-.29	.14	-.24	.04	-.20	-.15	.25	-.28	-.04	-.40	-.44
PCTCOL...	.23	.39	-.33	-.26	.15	.20	.31	.02	-.34	-.22	-.01	.22	-.25	-.04
PCTLUM...	.52	.39	-.29	-.15	-.03	.40	.39	.35	-.46	-.40	.18	.07	-.33	.29
PCTSTL...	-.18	-.08	.10	-.07	.28	-.26	.04	-.16	-.03	.15	-.30	.08	-.21	-.31
PCTOTH...	-.13	-.15	.09	-.06	.16	-.28	.05	-.02	-.07	.22	-.37	.03	-.37	-.26
PCTBEN...	.85	.95	-.72	-.62	.08	.79	.69	.06	-.52	-.75	.54	.33	-.49	.15
PCTBLK...	.11	.29	-.20	-.24	.37	.07	.28	-.04	-.26	-.18	-.05	.28	-.27	-.15
BENPB....	-.50	-.57	.75	.70	.21	-.84	-.73	.44	.62	.82	-.26	.01	.49	-.45
RELBEN...	.79	.79	-.67	-.56	.09	.85	.88	-.02	-.68	-.95	.31	.37	-.64	.50
PCTAPP...	-.48	-.41	.30	.27	.56	-.47	-.25	-.06	.16	.33	-.14	.27	.12	-.09
PCTMOV...	.25	.06	.09	.21	-.34	.15	.02	.11	-.07	-.21	-.09	.01	.09	.24
PHYSPP...	-.44	-.34	.52	.49	.20	-.51	-.73	.19	.71	.56	.21	.01	.82	-.50

Table 4. -- Continued

	PCTMIN	PCTCON	PCTMFG	PCTCOL	PCTLUM	PCTSTL	PCTOTH	PCTBEN	PCTBLK	BENPB	RELBEN	PCTAPP	PCTMOV	PHYSPP
MEDSCH...														
PCTHS4...														
PCTOLD...														
PCTPOV...														
UIPOV....														
PCTUNP...														
PCTURB...														
MEDFIN...														
PCTPA....														
PCTSS....														
PCTWCW...														
PCTAGR...														
PCTMIN...	1.00													
PCTCON...	.48	1.00												
PCTMFG...	-.36	-.48	1.00											
PCTCOL...	.63	.13	.04	1.00										
PCTLUM...	.05	.27	-.08	.04	1.00									
PCTSTL...	.02	-.30	.48	.27	-.19	1.00								
PCTOTH...	-.08	-.42	.66	.15	-.10	.65	1.00							
PCTBEN...	.37	.46	-.04	.46	.42	-.11	-.14	1.00						
PCTBLK...	.45	.03	.20	.79	-.01	.64	.30	.36	1.00					
BENPB....	-.20	-.65	.23	.04	-.28	.33	.42	-.53	.10	1.00				
RELBEN...	.55	.63	-.24	.35	.43	-.07	-.12	.80	.30	-.63	1.00			
PCTAPP...	-.26	-.49	.30	.05	-.17	.21	.19	-.28	.19	.33	-.30	1.00		
PCTMOV...	.14	.66	-.56	-.09	.25	-.37	-.43	.07	-.22	-.20	.19	-.40	1.00	
PHYSPP...	-.37	-.46	-.15	-.17	-.32	-.06	-.25	-.33	-.11	.42	-.57	.28	-.19	1.00

Notes: 1. There are 51 observations representing each of the 50 states and the District of Columbia. The data are weighted according to the number of persons in each State and the District of Columbia.

2. Under the assumption that $\rho = 0$, the probability of r exceeding .273 is .025. That is, values of r larger than .273 or smaller than -.273 are significantly nonzero at a 95 percent confidence level.

Definitions shown on following page.

Definitions:

PCTDIS - Percent of persons aged 16 to 64 with a partial or complete work disability.
PCTUNA - Percent of persons aged 16 to 64 with a complete work disability.
MEDSCH - Median school years completed for persons aged 25 and over.
PCTHS4 - Percent of persons aged 25 and over that completed 12 or more years of school.
PCTOLD - Percent of persons aged 16 to 64 that are aged 50 to 64.
PCTPOV - Percent of families in poverty.
UIPOV - Percent of unrelated individuals aged 14 and over in poverty.
PCTUNP - Percent of persons aged 16 and over that are unemployed.
PCTURB - Percent of persons that live in urbanized areas and in places of 2,500+ inhabitants outside urbanized areas.
MEDFIN - Median family income in 1969 less $9,500.
PCTPA - Percent of families receiving income from public assistance or welfare in 1969.
PCTSS - Percent of persons receiving income from Social Security or Railroad Retirement.
PCTWCW - Percent of workers aged 16 and over that are employed in white collar occupations.
PCTAGR - Percent of employed persons aged 16 and over that are employed in agriculture, forrestry or fisheries.
PCTMIN - Percent of employed persons aged 16 and over that are employed in mining.
PCTCON - Percent of employed persons aged 16 and over that are employed in construction.
PCTMFG - Percent of employed persons aged 16 and over that are employed in manufacturing.
PCTCOL - Percent of employed persons aged 16 and over that are employed in coal mining.
PCTLUM - Percent of employed persons aged 16 and over that are employed in lumber and wood product industries.
PCTSTL - Percent of employed persons aged 16 and over that are employed in blast furnace and steel working industries.
PCTOTH - Percent of employed persons aged 16 and over that are employed in other primary iron and steel industries.
PCTBEN - Number of Social Security disability beneficiaries per 100 persons.
PCTBLK - Number of Black Lung beneficiaries per 10,000 persons.
BENPB - Monthly Social Security benefit per beneficiary in 1970.
RELBEN - Average annual Social Security benefit in 1970 times 100 divided by the median family income in 1969.
PCTAPP - Percent of Social Security disability applications that are approved.
PCTMOV - Percent of residents that had moved from a different State since 1965.
PHYSPP - Number of physicians per 10,000 persons in 1969.

Table 5. -- Results From the Weighted Regressions

	Constant	MEDFIN	PCTPOV	PCTUNP	PCTPA	PCTSS	PCTWCW	PCTAGR	PCTMFG	PCTCOL
Dependent variable: The percent of persons with a partial or complete work disability										
Coefficient.....	11.8606	-.0005	.1321	.5332	*	*	*	*	.0246	*
"t" statistic....	(5.5)	(-2.8)	(2.6)	(5.6)	*	*	*	*	(1.6)	*
$\bar{R}^2$.....	.76									
Dependent variable: The percent of persons with a complete work disability										
Coefficient.....	2.9000	*	.1333	*	.2114	.1281	-.0456	-.0775	*	.1710
"t" statistic....	(3.3)	*	(6.7)	*	(5.0)	(3.5)	(-3.0)	(-3.5)	*	(3.3)
$\bar{R}^2$.....	.88									

* The variables are not significant.

Note: The above variables were selected on the basis of their theoretical relationship as well as their statistical significance.

THE COMPREHENSION FACTOR IN RANDOMIZED RESPONSE

Dennis M. O'Brien, University of Wisconsin - La Crosse
Robert S. Cochran, University of Wyoming

I. INTRODUCTION

Since the introduction of the randomized response technique for questioning interviewees on sensitive topics by Stanley L. Warner [4] in 1965, several modifications and extensions of the procedure have been presented. For two such extensions, see Greenberg, et al. [1] and Horvitz, et al. [2]. Oftentimes, the primary motivation for refinements has been to encourage further cooperation on the part of the potential respondent and thus provide more accurate information and make more precise estimates possible. In all applications of the technique, close adherence to the instructions and control over the implementation and mechanics is required. These later attempts to further assure anonymity sometimes carry with them a more complex set of instructions which the interviewee is expected to understand and then follow. Some investigations have been performed into the effects of truthfulness of the respondent on some of the questioning models. However, very little has ever been mentioned on the ability or the desire to follow instructions.

Comments contributed by respondents to a 'Consumer Opinion Survey' (see O'Brien, et al. [3]), where variations of the randomized response technique were used, indicate that there is reason to suspect less than complete comprehension and an unwillingness to follow instructions. Hence this paper will introduce a 'comprehension factor', which includes the idea of truthfulness as well as the interest in (and/or ability to) following instructions. Its effect on estimation and variance formulas will be shown for three randomized response models. Also considered will be the action taken by the 'non-comprehenders'.

II. INTRODUCTION OF THE COMPREHENSION FACTOR INTO TWO QUALITATIVE MODELS

The Warner related question procedure (see Warner [4]) requires that the respondent be given two statements of the form:

1) I am a member of Group A
2) I am not a member of Group A (1)

and a randomizing device. The respondent will use the randomizing device to determine to which statement he is to respond. His answer is then 'yes' or 'no'.

The Simmons unrelated question procedure (see Horvitz, et al. [2]) also uses a randomizing device, but the statements are now of the form:

1) I am a member of Group A
2) I am a member of Group B. (2)

In both models Group A is considered to be of a sensitive nature so that an individual when asked directly about his affiliation with that group may refuse to answer or may answer, but will give false information. Group B is of a non-sensitive nature and is assumed to generate no hesitancy in admitting membership. The goal of the Warner and Simmons procedures is to estimate π, the proportion of the population who are members of the sensitive Group A. For this paper it is assumed that π_Y, the proportion in Group B of the Simmons method, is known and hence only a single simple random sample of size n is needed. If π_Y is unknown, two samples are needed. For a discussion of this case, see Horvitz, et al. [2].

The maximum likelihood estimators and their variances for these two procedures are as follow:

Warner-

$$\hat{\pi}_W = \frac{P-1}{2P-1} + \frac{n_1}{(2P-1)n} \tag{3}$$

$$V(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2} \tag{4}$$

where P = the probability of the random device indicating Group A ($P \neq \frac{1}{2}$) and n_1 = the number of 'yes' responses.

Simmons-

$$\hat{\pi}_S = [\frac{n_1}{n} - (1-P)\pi_Y]/P \tag{5}$$

$$V(\hat{\pi}_S) = \frac{1}{P^2 n}[\pi P+\pi_Y(1-P)][(1-\pi)P+(1-\pi_Y)(1-P)]. \tag{6}$$

Under the assumptions of complete comprehension and truthfulness in responses, equal sample size, and equal probabilities of a Group A indication, $V(\hat{\pi}_S)$ is always less than $V(\hat{\pi}_W)$. However if, for any reason, a proportion of the respondents do not answer the question in the proper fashion then there is a possibility of circumstances developing where the Warner method may prove to have a lower mean-square-error. In the following all reasons for not answering in the proper fashion are grouped under the general heading of "comprehension".

To handle this concept the following additional parameters are introduced:

θ_W, θ_S for the levels of comprehension of the Warner and Simmons procedures, respectively (proportion of the sample that responds correctly and honestly);

θ_{YW}, θ_{YS} for the probability of responding with a 'yes' in the event of miscomprehending in the respective procedures.

For the present time, assume the values of these new parameters are unknown and thus cannot be allowed for in the estimators. Under this assumption the estimator expressions are unchanged but they are now biased and the variance expressions change. The bias and variance expressions are:

Warner-

$$\text{Bias}_W = (1-\theta_W)\left(\frac{P+\theta_{YW}-1}{2P-1} - \pi\right) \tag{7}$$

(Note that Bias_W is independent of the sample size n.)

$$V(\hat{\pi}_W) = \frac{1}{(2P-1)^2 n}\left[\theta_W\pi P+\theta_W(1-\pi)(1-P)+(1-\theta_W)\theta_{YW}\right]\cdot\left[\theta_W\pi(1-P)+\theta_W(1-\pi)P+(1-\theta_W)(1-\theta_{YW})\right]. \tag{8}$$

Simmons-

$$\text{Bias}_S = (1-\theta_S)\left(\frac{P\pi_Y+\theta_{YS}-\pi_Y}{P} - \pi\right) \tag{9}$$

(Note that Bias_S is independent of the sample size n.)

$$V(\hat{\pi}_S) = \frac{1}{P^2 n}\left[\theta_S\pi P+\theta_S\pi_Y(1-P)+(1-\theta_S)\theta_{YS}\right]\cdot\left[\theta_S(1-\pi)P+\theta_S(1-\pi_Y)(1-P)+(1-\theta_S)(1-\theta_{YS})\right]. \tag{10}$$

Comparing the two procedures on the basis of mean-square-errors, $MSE = V(\hat{\pi})+\text{Bias}^2$, under various parameter conditions it can be shown that situations exist where $MSE_W < MSE_S$. As an illustration, consider the situation where n = 500, P = .8, π = .3, π_Y = .1, θ_{YW} = .5, and θ_{YS} = .0. The following is observed as the comprehension factor increases:

$\theta_W = \theta_S$	MSE_W/MSE_S
.80	.62
.85	.77
.90	1.07
.95	1.68

Thus the comprehension levels and the action taken by the non-comprehenders should be considered when deciding which of these two models is to be implemented.

At this time, assume that pre-sampling or past experience has provided values for these new parameters. Allowing for the comprehension factor in the estimators makes them unbiased. The new estimators and their variances are:

Warner-

$$\tilde{\pi}_W = \frac{P-1}{2P-1} + \frac{n_1 - n(1-\theta_W)\theta_{YW}}{(2P-1)n\theta_W} \tag{11}$$

$$V(\tilde{\pi}_W) = \frac{1}{(2P-1)^2 n\theta_W^2}\left[\theta_W\pi P+\theta_W(1-\pi)(1-P)+(1-\theta_W)\theta_{YW}\right]\cdot\left[\theta_W\pi(1-P)+\theta_W(1-\pi)P+(1-\theta_W)(1-\theta_{YW})\right]. \tag{12}$$

Simmons-

$$\tilde{\pi}_S = \left[\frac{n_1-n(1-\theta_S)\theta_{YS}}{n\theta_S} - (1-P)\pi_Y\right]/P \tag{13}$$

$$V(\tilde{\pi}_S) = \frac{1}{P^2 n\theta_S^2}\left[\theta_S\pi P+\theta_S\pi_Y(1-P)+(1-\theta_S)\theta_{YS}\right]\cdot\left[\theta_S(1-\pi)P+\theta_S(1-\pi_Y)(1-P)+(1-\theta_S)(1-\theta_{YS})\right]. \tag{14}$$

Comparisons between the MSE of the 'standard' Warner and $V(\tilde{\pi}_W)$ reveal situations where the 'standard' is best, that is, where $MSE_W < V(\tilde{\pi}_W)$, as well as parameter combinations where the latter estimator is best. As an illustration, consider the case where n = 500, P = .8, π = .3, and θ_{YW} = .3. The following is observed as θ_W increases:

θ_W	$MSE_W/V(\tilde{\pi}_W)$
.7	1.12
.8	1.00
.9	.94

Similar situations occur for the 'standard' Simmons vs. the 'modified' Simmons. For example, consider the situation where n = 500, P = .8, θ_S = .8, and θ_{YS} = .5. The following is observed:

$\pi=\pi_Y$	$MSE_S/V(\tilde{\pi}_S)$
.1	15.29
.5	.67
.9	12.43

III. INTRODUCTION OF THE COMPREHENSION FACTOR INTO A QUANTITATIVE MODEL

The Greenberg quantitative model (see Greenberg, et al. [1]) uses the unrelated question randomized response procedure to obtain quantitative information on sensitive topics. A randomizing device is used to indicate the question to which the interviewee is to respond. The questions are of the form:

1) How many abortions have you had during your lifetime?
2) If a woman had to work full-time to make a living, how many children do you think she should have? (15)

Question 1) is considered the sensitive question, while 2) is considered the non-sensitive question. As for the Simmons model, the investigator may or may not know the parameter values for the responses to the non-sensitive question. In this case they are the mean and variance, denoted by (μ_Y, σ_Y^2). In this paper it is assumed they are known and thus a single sample of size n is needed. The object is to estimate the mean of the sensitive question response distribution, $\mu_{\underline{X}}$.

An unbiased estimator for $\mu_{\underline{X}}$ and its variance are given by:

$$\hat{\mu}_{\underline{X}} = [\bar{Z}-(1-P)\mu_Y]/P, \tag{16}$$

where $\bar{Z}$ is the sample response mean.

$$V(\hat{\mu}_{\underline{X}}) = \frac{1}{nP^2}\,[P\sigma_{\underline{X}}^2 + (1-P)\sigma_Y^2 + P(1-P)(\mu_{\underline{X}}-\mu_Y)^2], \tag{17}$$

where $\sigma_{\underline{X}}^2$ is the variance of the sensitive question response distribution.

Letting Θ be the unknown proportion comprehending and following all instructions and assuming all non-comprehenders respond as if answering the non-sensitive question, $V(\hat{\mu}_{\underline{X}})$ becomes

$$V(\hat{\mu}_{\underline{X}}) = \frac{1}{nP^2}\,[P\Theta\sigma_{\underline{X}}^2 + (1-P\Theta)\sigma_Y^2 + P\Theta(1-P\Theta)(\mu_{\underline{X}}-\mu_Y)^2]. \tag{18}$$

The estimator now has a bias of

$$\text{Bias}_G = (1-\Theta)(\mu_Y-\mu_{\underline{X}}). \tag{19}$$

The standard direct question estimator

$$\hat{\mu} = \bar{Z} \tag{20}$$

has a variance of

$$V(\hat{\mu}) = \sigma_{\underline{X}}^2/n \tag{21}$$

under complete truthfulness. Letting T be the probability of obtaining a truthful response in a direct question interview and assuming those not responding truthfully respond according to a distribution with mean and variance of μ_T and σ_T^2, the estimator has a bias and variance as follows:

$$\text{Bias}_D = (1-T)(\mu_T-\mu_{\underline{X}}) \tag{22}$$

$$V(\hat{\mu}) = \frac{1}{n}\,[T\sigma_{\underline{X}}^2 + (1-T)\sigma_T^2 + T(1-T)(\mu_{\underline{X}}-\mu_T)^2]. \tag{23}$$

Comparisons of the Greenberg MSE under varying degrees of comprehension and the direct MSE under varying levels of truthfulness reveal cases where the Greenberg procedure is best as well as cases where the direct question approach is best. As an illustration, consider the case where n = 500, P = .75, $\sigma_Y^2 = \sigma_T^2 = .5\sigma_{\underline{X}}^2$, $\mu_Y = \mu_T = .5\mu_{\underline{X}}$, $\sigma_{\underline{X}} = .1\mu_{\underline{X}}$, and T = .7. As Θ increases, the following is observed:

Θ	MSE_G/MSE_D
.6	1.78
.7	1.01
.8	.45
.9	.12

References

[1] Greenberg, B. G., Abernathy, J. R. and Horvitz, D. G., "Application of the Randomized Response Technique in Obtaining Quantitative Data," Proceedings of Social Statistics Section, ASA (Aug. 1969), 40-3.

[2] Horvitz, D. G., Shah, B. V. and Simmons, W. R., "The Unrelated Question Randomized Response Model," Proceedings of Social Statistics Section, ASA (1967), 65-72.

[3] O'Brien, Dennis M., Cochran, Robert S., Marquardt, Ray S. and Makens, James C., "Randomized Response vs. Direct Question in a Mail vs. Personal Interview Consumer Opinion Survey," College of Commerce and Industry Research Paper No. 85, University of Wyoming, July 1975, Laramie, Wyoming.

[4] Warner, S. L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," JASA, 60 (1965), 63-9.

A RANDOMIZED RESPONSE TECHNIQUE FOR INVESTIGATING SEVERAL SENSITIVE ATTRIBUTES

Ajit C. Tamhane, Northwestern University

1. INTRODUCTION AND SUMMARY

Most of the literature on randomized response (RR) techniques has been concerned with the study of a single sensitive attribute. However, very often, social researchers are interested in studying several sensitive attributes together. Therefore it is necessary to develop privacy preserving techniques which would allow statistical inference to be made concerning marginal as well as joint distributions of the attributes. Only recently attention of the survey statisticians has been focussed on this particular problem.

In his dissertation, Barksdale (1971) proposed and analyzed some RR techniques for investigating two sensitive dichotomous attributes. In particular, he considered a repeated (for each attribute) application of the Warner's original technique (see also Clickner and Iglewicz 1976 and Drane 1976), a repeated application of the Simmons' unrelated question technique (Greenberg et al. 1969) and a third technique which is as follows: The two statements concerning the two sensitive attributes are phrased so that a "Yes" response to one of the two statements would be nonstigmatizing. (E.g., the two statements might be "I have never smoked marijuana" and "I am an alcoholic.") The interviewer presents both the statements to the respondent on two occasions. On each occasion, the respondent picks one of the two statements at random, unknown to the interviewer, but according to some known probability (different for each occasion) and responds to it. This procedure maintains the privacy of the respondent and yet allows the researcher to compute the estimates of the marginal and bivariate probabilities of the attributes from the observed frequencies of "Yes-Yes," "Yes-No," "No-Yes," and "No-No" responses.

In a survey dealing with $t \geq 2$ sensitive attributes, a repeated application of any RR technique for a single attribute, such as the Warner's technique, involves t trials per respondent. If t is large then this procedure becomes tedious, costly and leads to degradation in cooperation on the part of respondents. Also the estimating equations involve all the joint probabilities which the researcher is not often interested in. On the other hand, the technique described in the previous paragraph can be easily extended to $t > 2$ case with the number of trials per respondent restricted to $r \leq t$ if the researcher's interest lies in only up to r-variate joint probabilities. Quite often, $r = 2$ will suffice for the purposes of the research.

In Section 2 of the present paper we extend the above technique (henceforth referred to as the multiple RR trials technique or the M-technique) to the case of $t > 2$ sensitive dichotomous attributes. But we restrict to only $r = 2$ trials per respondent to keep the algebra simple and also since $r = 2$ appears to be the most useful case from a practical viewpoint. The estimates derived by Barksdale do not satisfy the natural restrictions on the marginal and bivariate probabilities; also no procedure for testing independence between the attributes is provided in his work. We provide a correct statistical analysis of the extended technique and also give a test of pairwise independence for any set of pairs of attributes.

In Section 3 we carry out a numerical comparison of the multiple RR trials technique with some competing techniques in terms of the trace of the variance-covariance matrix of the estimator vector for the marginal and joint probabilities of the attributes. To make this a just comparison, it is necessary to keep fixed some measure of privacy afforded to the respondent. In Section 3.2 such a measure is defined which extends to $t > 1$, the corresponding notion for $t = 1$ due to Leysieffer and Warner (1976). No clear winner is indicated by the numerical comparisons which are made for the $t = 2$ case. But if the proportions in the population possessing the sensitive attributes are small (which is often the case) and the respondent jeopardy levels are moderate, i.e., not too high or low (which is also often the case) then the multiple RR trials technique appears to dominate. This technique has one drawback however, which is that it fails to attain certain low levels of respondent jeopardy. Still, in view of the practical advantages pointed out earlier, the multiple RR trials technique definitely merits a consideration in any survey dealing with several sensitive attributes.

2. MULTIPLE RR TRIALS TECHNIQUE

2.1 Notation and Description of Technique: Consider $t \geq 2$ dichotomous attributes $A_1, A_2, \ldots, A_t$; we shall assume that all the attributes are sensitive but obviously that need not be so. Let $\theta_{i_1 \ldots i_u}$ denote the unknown proportion of individuals in the target population which possesses the attributes $A_{i_1 \ldots i_u}$ $(1 \leq i_1 < \cdots < i_u \leq t,$ $1 \leq u \leq t)$. The researcher's interest lies in making statistical inference (estimation and hypothesis testing) concerning the θ's.

For employing the multiple RR trials technique, the statements must be phrased so that a "Yes" response to some statements would be nonstigmatizing whereas a "No" response to the others would be so. Without loss of generality, we shall assume that the first $s < t$ statements are phrased "I possess the attribute A_i" $(1 \leq i \leq s)$, a "No" response to each one of which would be nonstigmatizing; the remaining $t - s$ statements are phrased "I do not possess the attribute A_i" $(s + 1 \leq i \leq t)$, a "Yes" response to each one of which would be so. An appropriate choice of s would be $\tilde{=} t/2$. Let $\pi_{i_1 \ldots i_u}$ be defined in the same manner as $\theta_{i_1 \ldots i_u}$ but with respect to the modified attributes B_i which are either original A_i $(1 \leq i \leq s)$ or the complements of the A_i $(s + 1 \leq i \leq t)$. It is clear that the θ's can be obtained from the π's and vice versa and therefore we shall consider the equivalent problem of estimation of the π's.

As remarked in the previous section we shall assume that the researcher is interested only in the marginal and bivariate probabilities, i.e., $\pi_i\ (1 \leq i \leq t)$ and $\pi_{ij}\ (1 \leq i < j \leq t)$, respectively. Thus there are $t(t+1)/2$ unknown parameters to be estimated and only 2 trials may be performed per respondent. We now describe the technique.

A total sample of n individuals (which may be assumed to be a simple random sample drawn with replacement) is divided into $b \geq 1$ subsamples; the value of b will be specified in the following section. Let $n_1, n_2, \ldots, n_b$ be the subsample sizes with $\sum_{h=1}^{b} n_h = n$.

Each individual is presented all the t statements and asked to respond to one statement picked at random according to some randomizing device, but not reveal his choice of the statement to the interviewer. This procedure is repeated with another randomizing device and both the responses are recorded. Let $P_{hi}^{(\ell)}$ denote the (known) probability that an individual drawn from the hth subsample picks, on the ℓth trial, the ith statement $(1 \leq i \leq t)$; obviously we have $\sum_{i=1}^{t} P_{hi}^{(\ell)} = 1$ for $1 \leq h \leq b$ and $\ell = 1, 2$.

2.2 Estimation of the π's: Suppose that the responses are coded so that a score of $2^{\ell-1}$ is assigned to a "Yes" response on the ℓth trial and a score of 0 is assigned to a "No" response. Then the total score, say ν, completely identifies the individual's response. E.g., $\nu = 3$ corresponds to a "Yes-Yes" response, $\nu = 2$ corresponds to a "No-Yes) response etc. Let $\lambda_{h\nu}$ denote the probability of obtaining a score of ν for an individual drawn from the hth subsample. Then we have the following equations.

$$\begin{aligned}
\lambda_{h1} &= \sum_{i=1}^{t} P_{hi}^{(1)}(1 - P_{hi}^{(2)})\pi_i \\
&\quad - \sum_{i=1}^{t}\sum_{j=i+1}^{t} (P_{hi}^{(1)}P_{hj}^{(2)} + P_{hj}^{(1)}P_{hi}^{(2)})\pi_{ij} \\
\lambda_{h2} &= \sum_{i=1}^{t} P_{hi}^{(2)}(1 - P_{hi}^{(1)})\pi_i \\
&\quad - \sum_{i=1}^{t}\sum_{j=i+1}^{t} (P_{hi}^{(1)}P_{hj}^{(2)} + P_{hj}^{(1)}P_{hi}^{(2)})\pi_{ij} \\
\lambda_{h3} &= \sum_{i=1}^{t} P_{hi}^{(1)}P_{hi}^{(2)}\pi_i \\
&\quad + \sum_{i=1}^{t}\sum_{j=i+1}^{t} (P_{hi}^{(1)}P_{hj}^{(2)} + P_{hj}^{(1)}P_{hi}^{(2)})\pi_{ij} \\
\lambda_{h0} &= 1 - \lambda_{h1} - \lambda_{h2} - \lambda_{h3},
\end{aligned} \tag{2.1}$$

for $1 \leq h \leq b$. In the vector notation, if $\underset{\sim}{\lambda} = (\lambda_{11}, \lambda_{12}, \lambda_{13}, \ldots, \lambda_{b1}, \lambda_{b2}, \lambda_{b3})'$ and $\underset{\sim}{\pi} = (\pi_1, \ldots, \pi_t, \pi_{12}, \pi_{13}, \ldots, \pi_{t-1,t})'$ then (2.1) can be expressed compactly as

$$\underset{\sim}{\lambda} = \underset{\sim}{R}\,\underset{\sim}{\pi}, \tag{2.2}$$

where the elements of the matrix $\underset{\sim}{R}$ are given by the following equations: For $1 \leq h \leq b$ and $1 \leq i \leq t$ we have,

$$\begin{aligned}
\underset{\sim}{R}_{3h-2,i} &= P_{hi}^{(1)}(1 - P_{hi}^{(2)}), \\
\underset{\sim}{R}_{3h-1,i} &= P_{hi}^{(2)}(1 - P_{hi}^{(1)}),\ \underset{\sim}{R}_{3h,i} = P_{hi}^{(1)}P_{hi}^{(2)},
\end{aligned} \tag{2.3}$$

and for $1 \leq i < j \leq t$ if $k = it - i(i+1)/2 + j$ then we have

$$\begin{aligned}
\underset{\sim}{R}_{3h-2,k} &= -\,(P_{hi}^{(1)}P_{hj}^{(2)} + P_{hj}^{(1)}P_{hi}^{(2)}) \\
&= \underset{\sim}{R}_{3h-1,k} = -\,\underset{\sim}{R}_{3h,k}.
\end{aligned} \tag{2.4}$$

To find b, the total number of subsamples, necessary to estimate the t marginal probabilities $\{\pi_i\}$ and $\binom{t}{2}$ bivariate probabilities $\{\pi_{ij}\}$, consider an extreme case (and a most favorable one from the statistician's viewpoint) where the P-values can be chosen either equal to zero or one (which corresponds to the "direct response" case). By choosing $P_{hi}^{(1)} = 1$ and $P_{hj}^{(2)} = 1$ for different pairs (i, j) for different subsamples h, it is easy to see that all the parameters can be estimated using $\binom{t}{2}$ subsamples and no less number of subsamples would do. An extension of this argument shows that, even for general P-values, to estimate all the parameters, at least $\binom{t}{2}$ subsamples are required. In other words, by suitably choosing the P's, the matrix $\underset{\sim}{R}$ defined in (2.3) and (2.4) can be made to have a full column rank only if $b \geq \binom{t}{2}$. Let us then assume that $b \geq \binom{t}{2}$ and that $\underset{\sim}{R}$ is a full column rank matrix.

We propose to obtain the maximum likelihood estimator (MLE) of $\underset{\sim}{\pi}$ from the observed data $\{n_{h\nu}\}$ where $n_{h\nu}$ = the number of individuals from the hth subsample having a score of $\nu (0 \leq \nu \leq 3)$; $\sum_{\nu=0}^{3} n_{h\nu} = n_h (1 \leq h \leq b)$. The usual method of first obtaining the unrestricted MLE (UMLE) of $\underset{\sim}{\lambda}$ (i.e., the UMLE of $\lambda_{h\nu} = n_{h\nu}/n_h$ for $0 \leq \nu \leq 3$, $1 \leq h \leq b$) and then obtaining the UMLE of $\underset{\sim}{\pi}$ by "solving" (2.2) is not applicable for two reasons in the present context:

1. Matrix R can be chosen to be a square full rank matrix only for $t = 2$. For $t > 2$, in general, there is no unique solution in $\underset{\sim}{\pi}$ to (2.2).
2. Even in the case where the UMLE of $\underset{\sim}{\pi}$ can be obtained by the above method, the resulting estimator may not satisfy the natural restrictions on the π's namely that

$$\begin{aligned}
&0 \leq \pi_i \leq 1 \quad \forall i \quad \text{and,} \\
&\max(0, \pi_i + \pi_j - 1) \leq \pi_{ij} \leq \min(\pi_i, \pi_j)\ \forall (i, j).
\end{aligned} \tag{2.5}$$

From a theoretical viewpoint, the UMLE of $\underset{\sim}{\pi}$ may even be inadmissible as shown in the case of the Warner's technique for a single attribute by Singh (1976).

Therefore we must find the restricted MLE (RMLE) of $\underset{\sim}{\pi}$, say $\hat{\underset{\sim}{\pi}}$. We propose to obtain $\hat{\underset{\sim}{\pi}}$ directly by maximizing the likelihood function

$$L \propto \prod_{h=1}^{b}\prod_{\nu=0}^{3} (\lambda_{h\nu})^{n_{h\nu}} \tag{2.6}$$

subject to (2.5). In (2.6) the $\lambda_{h\nu}$ are given in terms of π by (2.1). Denote the restricted maximum of L by L*. The constraint set (2.5) is linear in the π's and the objective function $\log_e L$ can be easily checked to be concave in the π's. The resulting nonlinear programming (NLP) problem is thus well structured and can be solved quite economically on a computer using one of the commonly available algorithms.

2.3 Properties of $\hat{\underset{\sim}{\pi}}$: The RMLE $\hat{\underset{\sim}{\pi}}$ is biased in small samples but is asymptotically (as $n_h \to \infty \ \forall h$) unbiased. The asymptotic variance-covariance matrix of $\hat{\underset{\sim}{\pi}}$ (which is also the exact variance-covariance matrix of the UMLE of $\underset{\sim}{\pi}$) is given by the inverse of the information matrix $\underset{\sim}{\mathcal{I}}$; we give below an expression for the elements of the upper left $t \times t$ principal submatrix of $\underset{\sim}{\mathcal{I}}$: For $1 \leq i, j \leq t$ we have

$$\mathcal{I}_{ij} = -E\left\{\frac{\partial^2 \log L}{\partial \pi_i \partial \pi_j}\right\} = \sum_{h=1}^{b} n_h \sum_{\nu=0}^{3} \frac{1}{\lambda_{h\nu}} \left(\frac{\partial \lambda_{h\nu}}{\partial \pi_i}\right)\left(\frac{\partial \lambda_{h\nu}}{\partial \pi_j}\right).$$

The remaining elements of $\underset{\sim}{\mathcal{I}}$, which would involve $\partial \lambda_{h\nu}/\partial \pi_{ij}$ terms, can be obtained in an analogous manner. The various derivatives can be evaluated easily using (2.1).

For t = 2, the expressions for the asymptotic variances and covariances can be written down explicitly and they may be found in Barksdale (1971). Large sample hypothesis testing concerning the π's can be carried out using the expressions for the variances and covariances with $\underset{\sim}{\lambda}$ replaced by its RMLE $\hat{\underset{\sim}{\lambda}} = \underset{\sim}{R}\,\hat{\underset{\sim}{\pi}}$.

2.4 Test of Independence: First we note that testing pairwise independence between the original attributes, say A_i and A_j, is equivalent to testing pairwise independence between the corresponding modified attributes. In fact, if ρ_{ij} denotes the correlation between A_i and A_j and η_{ij} denotes the correlation between the corresponding modified attributes then $|\rho_{ij}| = |\eta_{ij}|$ for $1 \leq i < j \leq t$. Therefore we shall consider the problem of testing independence between pairs of modified attributes.

Suppose that it is desired to test the hypothesis $H_{\mathcal{J}}: \pi_{ij} = \pi_i \pi_j$ for all pairs (i, j) in a certain set $\mathcal{J}$. We can use the generalized likelihood ratio method to test this hypothesis as follows: Compute the maximum of the likelihood function L in (2.6) subject to the following constraints on the π's

$$0 \leq \pi_i \leq 1 \ \forall i,$$

$$\max(0, \pi_i + \pi_j - 1) \leq \pi_{ij} \leq \min(\pi_i, \pi_j) \ \forall (i, j) \notin \mathcal{J} \qquad (2.7)$$

$$\pi_{ij} = \pi_i \pi_j \qquad \forall (i, j) \in \mathcal{J}.$$

Denote the corresponding maximum of L by $L^*_{\mathcal{J}}$. Then under $H_{\mathcal{J}}$ asymptotically $-2 \log_e (L^*_{\mathcal{J}}/L^*)$ has a chi-square distribution with f degrees of freedom (d.f.), where f is the number of pairs in set $\mathcal{J}$.

2.5 Choice of the P's: For fixed h and ℓ, the $P_{hi}^{(\ell)}$ should be chosen as different from 1/t as possible. In fact, for large t, the length of the questionnaires can be cut down by choosing $P_{hi}^{(\ell)} = 0$ for different sets of statements for different subsamples. Assuming that the researcher is equally interested in all the attributes, it seems that, the P's should be chosen symmetrically as far as possible. For t = 2, such a symmetric choice is provided by $P_{11}^{(1)} + P_{11}^{(2)} = 1$; subject to this restriction, $P_{11}^{(1)}$ and $P_{11}^{(2)}$ may be chosen as far away from 1/2 as the researcher dares. Obviously the actual choice will depend on the average educational and social sophistication of the population. A pilot survey should be carried out to test different randomizing devices (different P's) as well as the questionnaire itself.

3. COMPARISON WITH SOME COMPETING TECHNIQUES

3.1 Brief Description of the Competing Techniques: We shall consider two techniques in competition with the M-technique developed above: a repeated application of the Warner's technique (W-technique) and a repeated application of the Simmons' unrelated question technique (S-technique).

In the W-technique t trials are performed per respondent. On the ith trial the interviewer presents the respondent with a pair of statements: "I possess the attribute A_i" and "I do not possess the attribute A_i." The respondent picks one of the two statements at random according to known probabilities P_i and $1 - P_i$ ($P_i \neq 1/2$) respectively, and without revealing his choice to the interviewer, responds to it. This procedure is repeated for $i = 1, 2, \ldots, t$. Suppose that the responses are coded so that a score of 2^{i-1} is assigned to a "Yes" response on the ith trial and a score of 0 is assigned to a "No" response ($1 \leq i \leq t$) and let ν denote the total score. Then $\nu (0 \leq \nu \leq 2^t - 1)$ completely identifies the individual's response. The π's can then be estimated from the observed frequencies $\{n_\nu\}$ where n_ν = the number of individuals in the sample having a score of ν; $\sum_{\nu=0}^{2^t-1} n_\nu = n$.

In the S-technique also t trials are performed per respondent. On the ith trial the interviewer presents the respondent with a pair of statements "I possess the attribute A_i" and "I possess the attribute Y_i" where Y_i is some unrelated and innocuous attribute. The respondent picks one of the two statements at random according to known probabilities P_i and $1 - P_i$ respectively, and without revealing his choice to the interviewer, responds to it. This procedure is repeated for $i = 1, 2, \ldots, t$. Again using the same scoring system as in the previous paragraph, the π's can be estimated from the observed frequencies $\{n_\nu\}$ if the fraction in the population possessing the attribute Y_i, say β_i, is known for $1 \leq i \leq t$.

3.2 A Measure of Respondent Jeopardy: Recently Leysieffer and Warner (1976) have developed a

measure of the jeopardy of respondent's privacy in the case of a single sensitive attribute. Here we shall extend their approach to the case of $t \geq 2$ sensitive attributes: Consider the 2^t mutually exclusive and collectively exhaustive groups into which the population is divided depending on the possession or nonpossession of different attributes and denote these groups by $A_1A_2\cdots A_t$, $A_1^cA_2\cdots A_t$, $\ldots$, $A_1^cA_2^c\cdots A_t^c$ where the notation is obvious. Consider, say, the group $A_1A_2\cdots A_t$. By using the Bayes' theorem in the same manner as Leysieffer and Warner (1976) it can be shown that a measure of information resulting from response ν in favor of $A_1A_2\cdots A_t$ against $(A_1A_2\cdots A_t)^c$ is given by

$$g(\nu;A_1A_2\cdots A_t) = P(\nu|A_1A_2\cdots A_t)/P(\nu|(A_1A_2\cdots A_t)^c). \qquad (3.1)$$

Thus the response ν can be regarded as jeopardizing with respect to the group $A_1A_2\cdots A_t$ (and not jeopardizing with respect $(A_1A_2\cdots A_t)^c$) if $g(\nu;A_1A_2\cdots A_t) > 1$ and not jeopardizing with respect to either $A_1A_2\cdots A_t$ or $(A_1A_2\cdots A_t)^c$ if $g(\nu;A_1A_2\cdots A_t) = 1$. Now to get a measure of the worst jeopardy of the privacy of an individual in group $A_1A_2\ldots A_t$ we define the jeopardy function for that group as

$$g(A_1A_2\cdots A_t) = \max_{\nu} g(\nu;A_1A_2\cdots A_t). \qquad (3.2)$$

The jeopardy functions for other groups can be defined in an identical manner.

The parameters of each RR technique should be chosen so that the jeopardy function values for different groups do not exceed some prespecified upper bounds. We note here that these jeopardy function values will depend in general on the unknown θ's (in contrast to the case of $t = 1$). Therefore some apriori guesses at the values of the θ's will be necessary to compute their values.

3.3 Jeopardy Functions for Competing Techniques: Using the definitions (3.1) and (3.2), we shall derive the expressions for the jeopardy functions associated with the W-, S- and the M-techniques for $t = 2$. Here we shall consider only the following special case of practical interest. (The general case with $t \geq 2$ is quite straightforward but algebraically messy and is hence omitted.) For the W-technique we take $P_1 = P_2 = P_W$ (say) where $P_W > 1/2$ without loss of generality. For the S-technique we take $P_1 = P_2 = P_S$ (say) and $\beta_1 = \beta_2 = \beta$ (say). For the M-technique we take $P_{11}^{(1)} = 1 - P_{11}^{(2)} = P_M$ (say) where $P_M > 1/2$ without loss of generality.

Define additional notation as follows: $Q_W = 1 - P_W$, $Q_S = 1 - P_S$, $Q_M = 1 - P_M$, $\gamma = 1 - \beta$ and $\theta_{12}^* = 1 - \theta_1 - \theta_2 + \theta_{12}$. Then the expressions for the jeopardy functions (using W, S and M to index the jeopardy functions for the W-, S- and the M-techniques respectively) are as follows. (The details of their derivations are given in an unabridged version of this paper available with the author.)

(i) W-technique:

$$g_W(A_1A_2) = P_W^2(1-\theta_{12})/\{P_WQ_W(1-\theta_{12}-\theta_{12}^*)+Q_W^2\theta_{12}^*\}$$

$$g_W(A_1^cA_2) = P_W^2(1-\theta_2+\theta_{12})/\{P_WQ_W(\theta_{12}+\theta_{12}^*)+Q_W^2(\theta_1-\theta_{12})\}$$

$$g_W(A_1A_2^c) = P_W^2(1-\theta_1+\theta_{12})/\{P_WQ_W(\theta_{12}+\theta_{12}^*)+Q_W^2(\theta_2-\theta_{12})\}$$

$$g_W(A_1^cA_2^c) = P_W^2(1-\theta_{12}^*)/\{P_WQ_W(1-\theta_{12}+\theta_{12}^*)+Q_W^2\theta_{12}\}.$$

(ii) S-technique:

$$g_S(A_1A_2) = (P_S+Q_S\beta)^2(1-\theta_{12})/\{Q_S\beta(P_S+Q_S\beta)(1-\theta_{12}-\theta_{12}^*)+Q_S^2\beta^2\theta_{12}^*\}$$

$$g_S(A_1^cA_2) = (P_S+Q_S\beta)(P_S+Q_S\gamma)(1-\theta_2+\theta_{12})/\{Q_S\gamma(P_S+Q_S\beta)\theta_{12}+Q_S^2\beta\gamma(\theta_1-\theta_{12})+Q_S\beta(P_S+Q_S\gamma)\theta_{12}^*\}$$

$$g_S(A_1A_2^c) = (P_S+Q_S\beta)(P_S+Q_S\gamma)(1-\theta_1+\theta_{12})/\{Q_S\gamma(P_S+Q_S\beta)\theta_{12}+Q_S^2\beta\gamma(\theta_2-\theta_{12})+Q_S\beta(P_S+Q_S\gamma)\theta_{12}^*\}$$

$$g_S(A_1^cA_2^c) = (P_S+Q_S\gamma)^2(1-\theta_{12}^*)/\{Q_S\gamma(P_S+Q_S\gamma)(1-\theta_{12}-\theta_{12}^*)+Q_S^2\gamma^2\theta_{12}^*\}.$$

(iii) M-technique:

$$g_M(A_1A_2) = P_M^2(1-\theta_{12})/Q_M^2\theta_{12}^*$$

$$g_M(A_1^cA_2) = (1-\theta_2+\theta_{12})/P_MQ_M(\theta_{12}+\theta_{12}^*)$$

$$g_M(A_1A_2^c) = (1-\theta_1+\theta_{12})/P_MQ_M(\theta_{12}+\theta_{12}^*)$$

$$g_M(A_1^cA_2^c) = P_M^2(1-\theta_{12}^*)/Q_M^2\theta_{12}.$$

3.4 Equating the Jeopardy Functions for the Competing Techniques: Our approach here will be to first equate the jeopardy functions for the four different groups for the competing techniques and obtain their equivalent parameter values, i.e., their P-values and the β-value for the S-technique. (Clearly the parameter values yielded by the four sets of equations will not in general be consistent. Therefore some criterion such as guaranteeing the lowest jeopardy level will be necessary in order to arrive at a unique parameter value for each technique.) The next step in our approach will be to compute for each technique a measure of its performance based on these parameter values. We have taken the measure of performance to be the trace of the variance-covariance matrix of the estimator vector. We note here that because of the special symmetric case that we are considering for each technique, no optimization in the sense of Leysieffer and Warner (1976) is possible.

First we equate the $g_W(\cdot)$'s with the respective $g_S(\cdot)$'s and we obtain that $P_S = 2P_W - 1$ and $\beta = 1/2$. Next we equate the $g_W(\cdot)$'s with the

respective $g_M(\cdot)$'s and solve the resulting quadratic equations for P_M in terms of $g_W(\cdot)$'s. We give below the condition that must be satisfied by $g_W(\cdot)$ in each case for the solution to be feasible (i.e., $1/2 \leq P_M \leq 1$) and the corresponding expression for P_M. For notational convenience we have defined the following quantities: $k_1 = g_M(A_1A_2)/(1 - \theta_{12})$, $k_2 = g_W(A_1^cA_2)/(1 - \theta_2 + \theta_{12})$, $k_3 = g_W(A_1A_2^c)/(1 - \theta_1 + \theta_{12})$ and $k_4 = g_M(A_1^cA_2^c)/(1 - \theta^*_{12})$. We have

$$g_W(A_1A_2) = g_M(A_1A_2) \Rightarrow P_M = (k_1\theta^*_{12} - \sqrt{k_1\theta^*_{12}})/(k_1\theta^*_{12} - 1)$$
$$\text{if } g_W(A_1A_2) \geq (1-\theta_{12})/\theta^*_{12}. \tag{3.3a}$$

$$g_W(A_1^cA_2) = g_M(A_1^cA_2) \Rightarrow P_M = [1+\{1-4/k_2(\theta_{12}+\theta^*_{12})\}^{1/2}]/2$$
$$\text{if } g_W(A_1^cA_2) \geq 4(1-\theta_2+\theta_{12})/(\theta_{12}+\theta^*_{12}). \tag{3.3b}$$

$$g_W(A_1A_2^c) = g_M(A_1A_2^c) \Rightarrow P_M = [1+\{1-4/k_3(\theta_{12}+\theta^*_{12})\}^{1/2}]/2$$
$$\text{if } g_W(A_1A_2^c) \geq 4(1-\theta_1+\theta_{12})/(\theta_{12}+\theta^*_{12}). \tag{3.3c}$$

$$g_W(A_1^cA_2^c) = g_M(A_1^cA_2^c) \Rightarrow P_M = (k_4\theta_{12} - \sqrt{k_4\theta_{12}})/(k_4\theta_{12} - 1)$$
$$\text{if } g_W(A_1^cA_2^c) \geq (1-\theta^*_{12})/\theta_{12}. \tag{3.3d}$$

It is only fair to point out that one drawback with the M-technique might be that it cannot match the W- and S-techniques at low levels of jeopardy. Also if unknown to the statistician, either $\theta_{12} = 0$ or $\theta^*_{12} = 0$ or both then at least one of the conditions on $g_W(\cdot)$ in (3.3) is certainly violated and there is no hope for matching the M-technique with the others in terms of the jeopardy values. In practice it is likely that θ_{12} (the proportion in the population possessing both the sensitive attributes A_1 and A_2) will be small whereas θ^*_{12} will be large. Hence it is likely that only the condition on $g_W(A_1^cA_2^c)$ in (3.3d) will be violated and it will not be possible to guarantee that $g_W(A_1^cA_2^c) = g_M(A_1^cA_2^c)$. However, this would be of no consequence since usually the upper limit on $g(A_1^cA_2^c)$ will be very large (even infinity) since $A_1^cA_2^c$ is a completely innocuous group.

Now the P_M-values given by (3.3a) - (3.3d) will in general be unequal. We follow the convention of guarding the individuals in the most sensitive group A_1A_2, i.e., controlling $g(A_1A_2)$ for each technique. Therefore we take the P_M-value given by (3.3a). Thus if the condition on $g_W(A_1A_2)$ in (3.3a) is satisfied then the corresponding P_M-value would be feasible and all the three techniques would be matched in terms of their jeopardy values for the A_1A_2 group.

3.5 Numerical Results: Define the trace inefficiency of a RR technique as the ratio of the trace of the (asymptotic) variance-covariance matrix of its estimates θ_1, θ_2, and θ_{12} to the corresponding quantity for the direct response technique when both the techniques use the same sample size n. This latter quantity is given by

$$\{\theta_1(1 - \theta_1) + \theta_2(1 - \theta_2) + \theta_{12}(1 - \theta_{12})\}/n.$$

The expressions for the traces of the variance-covariance matrices of the UMLE's of $\underset{\sim}{\theta}$ (which can be regarded as the asymptotic variance-covariance matrices of the RMLE's of $\underset{\sim}{\theta}$) for the W- and the S-technique are given respectively, in Clickner and Iglewicz (1976) and Barksdale (1971).

It can be checked that for the choice $P_S = 2P_W - 1$ and $\beta = 1/2$, the expressions for the variance-covariance matrices for the W- and the S-techniques are identical and therefore the two techniques are equivalent; this extends the corresponding result for $t = 1$ by Leysieffer and Warner (1976). Hence we only consider the comparison between the W- and M-techniques.

The values of P_W were obtained from Table 3 of Clickner and Iglewicz (1976) where they have computed them so that the W-technique attains selected levels of trace inefficiency (namely 1.25, 2.5, 5.0 and 10.0) for selected values of $\underset{\sim}{\theta}$. The corresponding values of P_M were computed from (3.3a) which guarantees that $g_W(A_1A_2) = g_M(A_1A_2)$ but does not in general guarantee the equality of the jeopardy levels for the other groups. Using these P_W and P_M values the trace inefficiencies for the two techniques were computed. The results of these computations are displayed in Table I. The values of P_M reported are rounded off in the third decimal place.

An inspection of the results reveals that if θ_1 and θ_2 are small (which would usually be the case for sensitive attributes) and P_W is in the range 0.7 - 0.8 (which are the values most frequently used in practice) then the M-technique indeed dominates the W-technique. However, for large values of θ_1 and θ_2 leading to small values of θ^*_{12} the situation is reversed and the M-technique has either very large values for trace inefficiency or in a few cases the M-technique is even nonexistent.

An explanation of this phenomenon is as follows: First, consider the variation with respect to θ^*_{12}. It is easy to check that for fixed P_W and θ_{12}, the P_M-value (as given by (3.3a)) decreases with θ^*_{12} which leads to high values of the trace inefficiency and in some instances even the nonexistence of the M-technique. Next consider the variation with respect to P_W. We note that, in general (i.e., except for the case $\theta_{12} + \theta^*_{12} = 1$), $P_M < 1$ even when $P_W = 1$ and therefore by a continuity argument we would expect the W-technique to dominate the M-technique for P_W-values in the neighborhood of 1. For fixed $\underset{\sim}{\theta}$, as P_W decreases, P_M decreases too. But for the intermediate values of P_W, it is possible for the M-technique to dominate the W-technique. As P_W decreases even further, P_M approaches 1/2 and therefore leads to very high values of the trace inefficiency for the M-technique.

No clear indication of the dependence of the trace inefficiency on ρ_{12} is evident in this table. It is known, however, that for the W- and the S-techniques, the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$ are not affected by the correlation; in fact, the corresponding formulae are the same as though these attributes were studied independently.

ACKNOWLEDGEMENT

The author wishes to thank Professor Michael Rubinovitch for a helpful discussion and Profes-

sor Robert Boruch for providing support for this research through Grant No. NIE-C-74-0115 from the National Institutes of Education.

REFERENCES

Barksdale, W. B. (1971), "New Randomized Response Techniques for Control of Nonsampling Errors in Surveys," unpublished Ph.D. thesis, Department of Biostatistics, University of North Carolina, Chapel Hill.

Clickner, R. P. and Iglewicz, B. (1976), "Warner's Randomized Response Technique: The Two Sensitive Questions Case," Report 5, Department of Statistics, Temple University; also presented at the Annual Meeting of the American Statistical Association at Boston.

Drane, W. (1976), "On the Theory of Randomized Responses to Two Sensitive Questions," Communications in Statistics-Theory and Methods, A5(6), 565-74.

Greenberg, B. G. Abul-Ela, A. A., Simmons, W. R., and Horvitz, D. G. (1969), "The Unrelated Question Randomized Response Model: Theoretical Framework," Journal of the American Statistical Association, 64, 520-29.

Leysieffer, F. W. and Warner, S. L. (1976), "Respondent Jeopardy and Optimal Designs in Randomized Response Models," Journal of the American Statistical Association, 71, 649-56.

Singh, J. (1976), "A Note on the Randomized Response Technique," paper presented at the Annual Meeting of the American Statistical Association at Boston.

Warner, S. L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," Journal of the American Statistical Association, 60, 63-69.

TABLE I

TRACE INEFFICIENCIES

θ_1	θ_2	θ_{12}	ρ_{12}	P_W	P_M	W-technique	M-technique
.05	.05	.0125	.2105	.988	.967	1.2449	1.6038
				.939	.911	2.5050	2.8568
				.878	.851	4.9937	4.7318
				.813	.789	9.9919	7.8812
.10	.05	.0250	.3059	.982	.953	1.2549	1.5824
				.918	.881	2.5022	2.7575
				.847	.812	5.0209	4.5591
				.781	.750	10.0041	7.5315
.20	.15	.0750	.3151	.963	.901	1.2785	1.5903
				.869	.798	2.4988	2.7253
				.789	.724	4.9772	4.5618
				.727	.669	9.9388	7.7277
.25	.05	.0375	.2649	.973	.912	1.2522	1.6233
				.888	.811	2.5019	2.8843
				.810	.738	4.9880	4.8655
				.745	.680	10.0223	8.3834
.25	.25	.0625	.0000	.962	.857	1.2511	1.6269
				.858	.729	2.5005	3.1724
				.777	.657	4.9902	6.1222
				.716	.606	10.0204	12.5720
.25	.25	.2500	1.0000	.953	.953	1.2478	1.2251
				.837	.837	2.4951	2.0434
				.756	.756	5.0077	3.2044
				.699	.699	10.0397	4.9160
.40	.05	.0250	.0468	.971	.872	1.2525	1.6838
				.882	.750	2.5027	3.2818
				.803	.675	4.9789	6.2459
				.738	.621	10.0390	12.7284
.55	.25	.1250	-.0580	.958	.784	1.2509	1.8043
				.848	.633	2.4987	5.2679
				.766	.561	4.9943	21.4589
				.706	.516	10.0363	294.7612
.75	.05	.0250	-.1325	.978	.784	1.2563	1.8443
				.904	.623	2.5031	6.6806
				.828	.538	4.9948	62.3038
				.760*	----	10.0363	------
.75	.70	.5250	.0000	.959	.676	1.2522	4.1165
				.850	.504	2.4924	7059.5656
				.766*	----	4.9977	------
				.705*	----	9.9647	------

Note: The M-technique does not exist for the starred P_W-values and the corresponding $\underset{\sim}{\theta}$ vectors.

ASSESSING THE EFFECTS OF MISSING DATA

John D. Hutcheson, Jr. and James E. Prather, Georgia State University

Problems resulting from incomplete data occur in almost every type of research, but survey research is especially prone to produce data sets within which some values for some subjects are missing. Many different methods for handling missing data have been proposed and employed. However, most commonly used procedures treat the missing data problem as "a disaster to be mitigated" rather than as a "pragmatic fact that may be investigated" (Cohen and Cohen, 1975: 288). In this paper, we discuss some of the problems associated with commonly used methods of handling missing data and illustrate the use of an alternative method, proposed by Jacob Cohen (1968: 438), that treats the fact that there are missing observations in a manner which permits the researcher to identify how missing observations could affect analysis and interpretation. Missing data are viewed, from the perspective explained below, as a specification problem rather than as a technical inconvenience.

Commonly Used Methods of Handling Missing Data

Most commonly used methods of handling missing data assume that missing observations occur randomly. When employing such procedures in survey research, the implicit assumption is that the fact that some respondents refuse to answer or are unable to respond to some questions is not related to any of the other items included in the analysis of the data. A similar assumption is used when it is assumed that refusals to participate as a respondent in surveys occur randomly. Nonrespondents are assumed to be similar to respondents or to differ only in ways unrelated to the content of the survey instrument. Hesseldenz (1976) has treated this type of nonresponse as a specification problem and has shown how the nature of nonresponse bias may be examined to enhance analysis and interpretation. The perspective and approaches employed below are similar. However, the problem addressed here is the effect of missing data for specific items rather than the effect of missing data for entire cases (respondents). Since most surveys include attitudinal items and/or questions related to personal characteristics, the respondent, while permitting the interview, may refuse or be unable to respond to specific questions. The assumption that this occurs in a manner which is not related to other variables included in the analysis of the data merits examination.

While assuming that missing observations occur randomly (Hertel, 1976; Gleason and Staelin, 1975; and Press and Scott, 1976), most commonly used methods for handling missing data can result in undesirable effects, in addition to incorporating this uninvestigated assumption. Case-wise (also referred to as list-wise) and pair-wise deletion of missing data, often employed in analyses using software packages such as the Statistical Package for the Social Sciences (SPSS), redefine the original sample to include what could be an unrepresentative subsample of the population and ignore factors that could be important in interpretation.

Many researchers have attempted to "plug" missing observations by substituting sample means for missing data. This procedure, again, assumes that missing observations occur randomly. Additionally, using the mean substitution method reduces the total variance observed and results in conservative estimates of association. These procedures fail to incorporate the informational value of missing data in a manner which allows the researcher to determine if analyses are biased (misspecified), and may, therefore, result in misinterpretation.

It has been argued that the "best one can do is select a missing data routine which does not increase biases already in the available data" (Hertel, 1976: 460). We contend, however, that the methods suggested by Cohen (1968: 438), and described, illustrated and expanded below, permit the researcher to assess the effects of missing data in a specific analysis and consequently minimize the possibility of undetected bias and misinterpretation. This, we believe, is more desirable than merely insuring that the existing bias is not increased.

An Alternate Method

The method described here allows the researcher to estimate, first, if missing data are systematically associated with substantive variables in a given analysis. If systematic relationships do exist, that is, if missing observations are related to other variables in the analysis, the researcher may then assess the nature of such bias and incorporate this additional information into the interpretation of the analysis.

The method which Cohen (1968: 438) has suggested and which is described in more detail in Cohen and Cohen (1975: 265-290) involves the substitution of sample means for missing observations. In contrast to the mean substitution method, however, this procedure concurrently employs the creation of dummy variables for every variable in which means have been substituted. For example, if a sample mean for education is 11.8 years and 50 observations of the 250 in the sample are missing, the value 11.8 would be used to "plug" the missing observations and a dummy variable, "missing education," would be created assigning a value of "0" to each actual response to the item and a value of "1" to each missing observation.

If education is an independent variable in a given analysis, a regression model may be used to identify systematic relationships of the actual responses to the education item and the missing education index with the dependent variable. If the analysis results in a significant relationship between the missing education index and the dependent variable, in the presence of the actual education variable, missing observations in education have not occurred randomly and the researcher may assess the nature and consequences of the resulting bias.

If, on the other hand, education is a dependent variable, with a limited number of categories, the researcher could create a missing education category to employ along with the other education categories so that discriminant analysis could be used to identify systematic differences between those subjects who responded to the education item and those who did not. Such procedures utilize all available information, including the fact that some respondents provide valid information and some do

not. This allows the researcher to assess the nature of the bias that may be introduced by missing data.

An Example

Participation of eligible voters in the electoral process has been considered a critical feature of traditional democratic theory. Consequently, many scholars have sought to explain why some eligible citizens vote and others do not. Much of what we know about voter participation is based upon survey data.[1] Research employing survey methods, for example, has found that participation is associated with education, income and other measures of social status (Verba and Nie, 1972: 125-137). Ben-Sira (1977) has synthesized a large body of research in formulating an explanation for the positive relationship between social status and voting. In general, it is thought that higher status is associated with greater resources (education, income) and that "the greater one's resources (or personal power potential) the greater one's striving for realization of this potential through achievement of a higher level of power" (Ben-Sira, 1977: 1970). The higher status person, then, has higher levels of political interest (Berelson, Lazarsfeld and McPhee, 1954: 25), political efficacy (Campbell, Converse, Miller and Stokes, 1964: 253) and a greater awareness "of the impact of government on the individual" (Ben-Sira, 1977: 1970).

A survey of over 7,000 residents (a 1% sampling) of Atlanta and Fulton County (part of suburban Atlanta), Georgia, conducted in mid 1976, provides an opportunity to explore some of the hypotheses posited above. This analysis, then, provides the context within which the consequences of employing differing procedures for handling missing data are illustrated. The 1976 Atlanta/Fulton County survey provides a data set with which to demonstrate these consequences when commonly used missing data procedures are most justifiable -- when the sample size is unusually large. The impacts of commonly used missing data routines, in most instances, will decrease as the size of the sample increases. Thus, if negative consequences are apparent in the analysis of the Atlanta/Fulton County survey data, they are likely to be quite severe when these routines are used with smaller data sets.

In this example, the dependent variable is voting, having voted in the past 5 years (1) or not (0). In the first analysis presented here, regression is used to identify bias attributable to missing observations in the independent variables. Explanation of voter participation is sought in terms of socio-economic variables (race, family income, education and age) and political attitudes (political efficacy and interest and governmental salience). In accord with the above theory, education, family income and political efficacy and interest are hypothesized to have positive effects on voter participation, as is whether or not the respondent felt that government had a "good deal" of impact on him/her, as an individual (governmental salience). Race and age are included in the model as control variables. Non-white was coded "0" and White was coded "1', since previous research suggests a positive association between "being White" and voter participation (Campbell, Converse, Miller and Stokes, 1964: 150). Since respondents in the survey were as young as 18 years of age, and because voting and non-voting were operationally defined to include the 5 years prior to the survey, the necessity to control for age is apparent.[2]

The voter participation model was estimated using four different specifications for missing data (see Table 1). The case-wise (or list-wise) deletion routine eliminates any case in which there are missing observations for any of the variables included in the model. The pair-wise deletion technique causes any case with a missing observation for a particular variable to be deleted from calculations involving that variable only. The mean substitution specification, as described earlier, "plugs" missing observations for variables with sample means. The fourth missing data specification is the mean substitution and dummy variable method described in the previous section.[3]

Table 1 provides the regression estimators and standard errors for each of the four specifications of the voter participation model. From an examination of the missing data variables in the fourth specification, it is clear that missing observations in the education, age, political efficacy and governmental salience variables are systematically related to voting. Only one of the three missing family income variables is significantly related to voter participation; while refusing to provide income information and missing income (no reason given) do not occur systematically, not being able to provide family income data ("don't know family income") is systematically related to voting.[4]

While the estimators and standard errors in the four specifications are fairly consistent, the lower R^2's for the case-wise (list-wise) deletion and the mean substitution specifications reflect the decrease in variance attributable to loss of cases and degrees of freedom in the case-wise (list-wise) deletion specification and to loss of variance in the mean substitution specification. Thus, even in analyses of a survey data set much larger than most, it appears that deletion and mean substitution methods do have some impact on the estimation of the model.

In Table 1, the use of the dummy variables in the fourth specification demonstrates that persons not responding to the education item, answering that they do not know their family income, not answering the governmental salience question or not responding to one or more of the items in the political efficacy scale are less likely to have voted than persons responding to these items. In general, the analysis suggests that persons not responding to political-attitude items are less likely to participate in the electoral process through voting. With respect to the political efficacy scale, the fact that a respondent did not respond to one or more of the efficacy items is more meaningful in the analysis than any set of valid responses. Respondents not reporting their age were more likely to have voted than respondents answering the age question.

When comparing the estimators for each set of valid and missing variables in the fourth specification, the analysis is consistent with the suppositions that older persons are less likely to respond to age items, that persons with less education and whose families have less income are less likely to respond to education and income questions, and that people who feel that government has little

TABLE 1

MODELS OF VOTER PARTICIPATION USING DIFFERENT SPECIFICATIONS FOR MISSING DATA

VARIABLES	CASE-WISE DELETION		PAIR-WISE DELETION		MEAN SUBSTITUTION		MEAN SUBSTITUTION AND DUMMY VARIABLES	
	ESTIMATOR	STANDARD ERROR OF ESTIMATOR	ESTIMATOR	STANDARD ERROR OF ESTIMATOR	ESTIMATOR	STANDARD ERROR OF ESTIMATOR	ESTIMATOR	STANDARD ERROR OF ESTIMATOR
EDUCATION (YEARS)	.033	.0025	.038	.0025	.039	.0017	.036	.0017
AGE (10 YEAR UNITS)	.032	.0048	.041	.0046	.038	.0032	.041	.0032
FAMILY INCOME ($10,000 UNITS)	.018	.0055	.029	.0059	.021	.0050	.023	.0050
RACE (WHITE)	-.039	.0160	-.051	.0160	-.032	.0110	-.043	.0110
POLITICAL EFFICACY SCALE	-.0028	.0160	-.0021	.0048	-.0016	.0036	-.001	.0036
POLITICAL INTEREST SCALE	.004	.00081	.0061	.00089	.0066	.00062	.0058	.00061
GOVERNMENTAL SALIENCE	.029	.0150	.040	.0150	.043	.0110	.044	.0100
CONSTANT	.21		.047		.0071		.098	
MISSING DATA INDICATORS:								
MISSING EDUCATION							-.190	.0240
MISSING AGE							.130	.0300
MISSING FAMILY INCOME							-.045	.0190
'DON'T KNOW' FAMILY INCOME							-.097	.0110
REFUSED TO GIVE FAMILY INCOME							.018	.0160
MISSING RACE							.025	.0250
MISSING POLITICAL EFFICACY							-.058	.0130
MISSING POLITICAL INTEREST							-.028	.0180
MISSING GOVERNMENTAL SALIENCE							-.090	.0190
R^2	.13		.16		.14		.17	
STANDARD ERROR	.36		.41		.41		.40	
N	2646		3138		7018		7018	

impact on them are less likely to respond to the governmental salience question. It further suggests that the estimators for the valid education, age, income, political efficacy and governmental salience variables are rendered slightly more conservative by the occurrence of missing observations. Thus the analysis using the mean substitution and dummy variable specification permits the researcher to assess the systematic effect introduced by missing observations.

While the use of the mean substitution and dummy variable specification provides more information for use in interpretation, and while the bias attributable to missing observations is discernible, such bias, in the analysis presented here, does not appear to be such that it would lead to misinterpretation when commonly used missing data routines are employed. All of the specifications included in Table 1 support the hypothesis that higher socio-economic status persons are more likely to vote. While political interest and feeling that government affects one as an individual (governmental salience) are positively related to voting, political efficacy seems to have little, if any, effect on voting. "Being White," when other variables are controlled, has a negative effect on voting. This finding is inconsistent with some prior voting behavior studies, but is supported by previous research in Atlanta, a city where Black candidates are major actors in electoral politics (Collins, 1977). The systematic bias introduced by missing observations in this model seems to have moderate impact; some bias is apparent and such bias could be much more severe when analyzing smaller data sets or data sets with more missing observations. In such instances, bias may lead to misinterpretation if the assumption of random occurrence of missing observations is not investigated.

Even though there were few missing observations (.8%) in the dependent variable in this example, the use of discriminant analysis will illustrate the utility of the mean substitution and dummy variable procedure in estimating the effects of missing observations in dependent variables. Discriminant analyses were performed on two specifications of the voter participation model; the first employed voting, non-voting and missing categories and the mean substitution procedure to "plug" missing observations in the discriminating variables. In the second discriminant analysis, the same categories were used, but the mean substitution and dummy variable procedure was used to create a missing index for each discriminating variable and these indices were employed along with the valid discriminating variables.

In the first analysis, using the mean substitution procedure, 57.03% of the "grouped" cases were correctly classified. When employing the mean substitution and dummy variable specification, 65.5% of the "grouped" cases were correctly classified. It is clear, then, that including indices of missing observations provides additional substantive information. Again, this procedure, unlike most commonly used procedures that attempt to "get rid" of missing data,[5] allows the researcher to determine if the fact that missing data occurred has substantive import. In this case, it obviously does.

Additionally, employing commonly used missing data routines without investigating the random occurrence assumption can increase rather than decrease existing bias attributable to missing observations. Table 2 presents the prediction table resulting from the discriminant analysis employing the mean substitution and dummy variable specification. Note that missing observations did not occur randomly; the analysis shows that the missing category could be systematically predicted along with the substantive categories. Furthermore, this analysis suggests that respondents that did not answer the voting question were more likely, if they had responded, to have reported that they did not vote than to have reported that they did vote. Seventy-three percent of the subjects responding to the voting item reported that they had voted ($\bar{X}$=.73). Thus, the analysis suggests that had the mean substitution procedure been employed to "plug" or "get rid" of missing observations in the voting participation variable, bias would have been exacerbated rather than mitigated. If the missing observations in the voting variable were to be "plugged," the value assigned them should reflect the greater likelihood of the respondents having reported that they did not vote. If cases with missing observations in the voting variable were to be allocated either to voting or non-voting categories, it would be appropriate, on the basis of the variables employed here, to allocate a larger proportion of these cases to the non-voting category.

Summary

In the example above, it is shown that the assumption of random occurrence of missing observations may be investigated in order to determine if bias is introduced by missing data. If bias is identified, the procedures explained above can help the researcher in understanding the nature of such bias, thereby enhancing interpretation of the analysis. While bias attributable to the occurrence of missing observations may or may not influence analysis in a manner which affects interpretation, we contend that the mean substitution

TABLE 2

PREDICTION RESULTS OF DISCRIMINANT ANALYSIS USING VOTED, DID NOT VOTE AND "MISSING" VOTING CATEGORIES AND EMPLOYING VALID AND MISSING INDICES AS DISCRIMINATING VARIABLES

ACTUAL GROUP	PREDICTED GROUP			N
	VOTED	DID NOT VOTE	"MISSING" VOTING	
VOTED	[64.1%]	31.0%	4.8%	5136
DID NOT VOTE	23.4%	[69.7%]	6.9%	1882
"MISSING" VOTING	18.0%	27.9%	[54.1%]	61

% OF "GROUPED" CASES CORRECTLY CLASSIFIED: 65.5%

and dummy variable procedure explained above provides a relatively simple method for clarifying the effects of missing data and incorporating additional information into interpretation and analysis. Missing data, here, are viewed as additional information to be analyzed and understood, rather than discarded.

Commonly used missing data routines employed without investigating the random occurrence assumption have several disadvantages. These routines not only "get rid" of information that may be useful in the analysis of survey data, but they also, as has been demonstrated above, may obscure or even increase existing biases. The mean substitution and dummy variable procedure permits the researcher to use all available information in an attempt to fully understand the effects of both valid and missing observations.

Notes

1. See for examples: Berelson, Lazarsfeld and McPhee (1954); Campbell, Converse, Miller and Stokes (1964); and Verba and Nie (1972).

2. Education is defined as number of years of formal schooling; family income is defined as total family income, before taxes and other deductions, for the calendar year 1975. The political efficacy and interest variables are three-item scales adapted from Verba and Nie (1972: 367-370).

3. The numbers of missing observations for the independent and control variables in the model are as follows: Missing Education (342); Missing Age (224); Missing Race (266); Missing Political Efficacy (1,469); Missing Political Interest (562); Missing Governmental Salience (635); Missing Family Income (524); Refused to Give Family Income (772); "Don't Know" Family Income (1,963).

4. Standard errors of estimators in regressions employing dummy variables should be interpreted cautiously since there is a tendency for the standard errors to be artifically deflated (Matloff, 1977). Therefore, we have interpreted the stability of the estimators in this model conservatively.

5. For examples of such procedures in discriminant analysis, see Chan, Gilman and Dunn (1976).

References

Ben-Sira, Zeev (1977) "A Facet Theoretical Approach to Voting Behavior," Quality and Quantity 11 (June): 167-188.

Berelson, B.R., P.F. Lazarsfeld and W.N. McPhee (1954) Voting. Chicago: University of Chicago Press.

Campbell, A., P.F. Converse, W.E. Miller and D. E. Stokes (1964) The American Voter. New York: John Wiley and Sons.

Chan, L.S., J.A. Gilman and O.J. Dunn (1976) "Alternative Approaches to Missing Data in Discriminant Analysis." Journal of the American Statistical Association 71 (December): 842-844.

Cohen, Jacob (1968) "Multiple Regression as a General Data-Analytic System," Psychological Bulletin 70 (December): 426-443.

Cohen, Jacob and Patricia Cohen (1975) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: John Wiley and Sons.

Collins, William P. (1977) "Race as a Salient Factor in Non-Partisan Elections." Unpublished Manuscript, Department of Political Science, Georgia State University.

Gleason, Terry C. and Richard Staelin (1975) "A Proposal for Handling Missing Data." Psychometrika 40 (June): 229-252.

Hertel, Bradley R. (1976) "Minimizing Error Variance Introduced by Missing Data Routines in Survey Analysis." Sociological Methods and Research 4 (May): 459-474.

Hesseldenz, Jon S. (1976) "Determining Validity and Identifying Nonresponse Bias in a Survey Requesting Income Data." Research in Higher Education 5 (April): 179-191.

Matloff, N.S. (1977) "Realistic Standard Errors of Estimated Regression Coefficients." Paper presented at the Western Meeting of the Institute of Mathematical Statistics and the American Statistical Association, June 22-24, Stanford California.

Press, S.J. and A.J. Scott (1976) "Missing Variables in Bayesian Regression, II." Journal of the American Statistical Association 71 (June): 366-369.

Verba, Sidney and Norman H. Nie (1972) Participation in America. New York: Harper and Row.

TWO DIRECT APPROACHES TO SURVEY NONRESPONSE: ESTIMATING A PROPORTION WITH CALLBACKS AND ALLOCATING EFFORT TO RAISE RESPONSE RATE

Charles H. Proctor, North Carolina State University

1. Introduction

The impetus for the work reported here was a problem of nonresponse in a survey of North Carolina dairy farms. That problem will be described and treated briefly in the last section. First, an outline and review is presented of approaches to nonresponse that are so excellently surveyed in the sampling textbooks, particularly Cochran [4] and Kish [12], as well as Hansen, Hurwitz, Madow [8], Sukhatme and Sukhatme [18] and Deming [6]. Also the somewhat specialized sample survey treatments of nonresponse are related to more conventional methods in statistics. In particular, maximum likelihood scoring is applied to data from repeated callbacks in order to estimate an underlying population proportion.

2. Some Passive or Post Hoc Approaches to Nonresponse

Data that are lost or misplaced in the office, that were not collected because of equipment failures or that are missing for any reasons that are clearly part of a causal nexus almost completely disjoint from that which sets the levels of the variables of interest in the survey, can usually be handled as a case of simple random subsampling from the initial sample (see Rubin [17] for conditions that allow this). When the initial design specifies simple random sampling this approach is easily implemented by reducing sample size to the number of respondents. When, however, unequal probabilities or other complex design features were used it is likely to be tedious to make adjustments. Some form of "hot deck" imputation as described by Chapman [3], may be useful in this case. "Hot deck" imputation requires location of a respondent with inclusion probability and other auxiliary characteristics similar to those of the nonrespondent and attribution of that respondent's data to the missing case. "Cold deck" imputation also refers to use of a similar case's data, but from an earlier survey.

In contrast to the disjoint causal nexus case are instances where data are missing whenever the variable of interest falls outside a critical range. Light bulbs, for example, are burned until they fail or until 100 days go by, at which time the test may be discontinued. Such data are taken to be drawn from a truncated distribution. A similar, but as Kendall and Stuart point out [11, p. 522 ff], a theoretically distinct case is that of censored data. The criterion for excluding data here is the relative standing of the observations among others in the sample. The practice of discarding the largest or the smallest observations has been found to be positively beneficial in some cases. It is perhaps better in gaining approval of the method to call it "trimming" rather than "throwing data away." These topics of truncated distributions, censoring, discarding outliers, trimming or making use of what are called "robust" estimators are, of course, too extensive to review, but it may be helpful to recognize their kinship to methods more in the traditions of sample survey work, to be discussed presently.

There is an approach related to truncation that can sometimes resolve cases of nonresponse as well as problems of outliers. It may happen, upon scanning what can be learned of a case of either nonresponse or "outlying," that the unit should not have been included in the population of interest in the first place. Such an observation can then be declared a blank.[1] There may arise the problem of knowing the size of the surveyed population since the original frame size was apparently too large. However, if one is interested in estimating a mean or other ratio, there may be no critical need of exact knowledge of the population size. The solution then is simply to declare the missing data case as outside the scope of interest.

As an example where this method may be employed one thinks of a legislator polling his constituients. Those who do not reply can be declared to be insufficiently politically active to have ideas that are of interest. In surveys of product preferences in market research the investigator may simply not care what household members may think who are not sufficiently motivated by the "free offer" to return the survey forms. The method may appear to be a somehow shoddy practice, perhaps because it is so inexpensive, but I feel that it could be used more widely. At any rate it is often done surreptitiously when, if it enjoyed a bit higher degree of respectability, it may be more often acknowledged and this would help to make survey reporting more complete and honest.

3. Some Indirect Adjustment Methods

Now we arrive at the more common cases of nonresponse, in which a tie can be identified between the process producing the missing data and that behind the variable of interest. There are a group of methods used in household surveys that call for the collection of additional data directly relevant to the causes of nonresponse, mainly being "not at home," and then make corrections based on this auxiliary information. We will not attempt an extensive review of these, as already appears in Cochran [3, pp. 371-374], but only describe three of them briefly to see how they compare operationally to other approaches that might be used.

In the Politz-Simmons [15] method persons are called upon once only, but if found at home are queried as to whether they were at home: "At this time yesterday?" and four more times are asked: "And the day before?" Data from those persons who report being away more often are expanded to that extent in the tabulations to account for calls that found no one at home. In Bartholomew's [1] method, where no one is found at home the interviewer goes to neighboring houses and apartments to determine a time to return which will maximize the probability of finding someone there. He then calls back only once more. Initially successful interviews are treated as from one stratum and successful callback interviews as from a second stratum. Different expansion factors are used for the two strata. Finally, in Kish and Hess' [13]

method, addresses that in earlier surveys yielded not-at-homes are added to those in the current selected sample in such numbers that the resulting completed interviews need not be differently weighted.

All three of these methods depend on well-trained interviewers carrying out their instructions carefully. All three are most useful when conducted by a fairly large-scale survey organization. When used responsibly, any one of the three methods can be very effective in reducing bias caused by not-at-homes.

4. Some Direct Approaches

Moving now from the adjustment methods brings us to approaches that depend on the controlled use of a variety of techniques for collecting data from mobile or reluctant respondents. The simplest approach is perhaps to "throw money at the problem," a substantial payment to the respondent for a completed schedule is worthy of serious consideration. Alternatives in survey methods that might affect rates of nonresponse include:

1) Pre-contact publicity using letter of introduction, media publicity, various sponsorships, local clearances, professional certifications, etc.

2) Telephone versus mail versus personal visit.

3) Use of interviewers whose sex, ethnic membership, social standing, etc. are different from or the same as respondent's.

4) Interviewers being experienced or not or local or not.

5) The survey instrument is loosely focused with a check list, it is a schedule or detailed outline, or it is a tightly structured though naturally worded questionnaire.

6) There is a lengthy explanation of survey objectives, guarantees of anonymity and confidentiality or such are minimal.

7) Randomized response or unrelated question methods may be used for sensitive topics.

8) A legal obligation to respond may be invoked.

Four styles of using such varying efforts in a systematic way in combatting nonresponse may be distinguished:

1) Make repeated calls (i.e., callbacks), mailings or telephone dialings using the same approach each time. In extreme cases one can reduce nonresponse in this way, a little bit at a time, almost "forever." In other cases the no-further-return plateau arrives quickly.

2) Make an initial attempt using a relatively inexpensive approach to a fairly large sample and then subsample the nonrespondents for applying a fairly elaborate approach that can almost guarantee response from those in the subsample.

3) Consider a continuum of approaches along a proportion nonresponse by effort curve, locate the optimum level of effort to devote to nonresponse relative to other survey expenses, and carry out that one level of effort for each and every selection in the sample.

4) Use a graduated series of approaches until the respondent is induced to respond.

Style 2, the use of a follow-up subsample as developed by Hansen and Hurwitz [8], has been widely and successfully used. It has more recently been given a Bayesian formulation by Erikson [7]. Whenever there is any appreciable number of hard core nonrespondents, this tends to upset the method. A relatively minor additional point is the possible presence of differing measurement biases for the initial as compared to the follow-up interviews. This drawback of differing measurement biases probably becomes more aggravated under Style 4, the "escalation" approach. Here, however, there enter also considerations of fair treatment of respondents. It seems to me unjust either to reward recalcitrant respondents with high payments or on the other hand to apply legal compulsion or punishment only to the reluctant ones. Because of these problems with Style 4 and the already available material on Style 2, further consideration is given only to methods of Style 2 and Style 3, the repeated calls case and the optimal choice of uniform level of effort to reduce nonresponse.

5. Maximum Likelihood Estimation of a Proportion in a Survey With Callbacks

The relative performance of differing numbers of calls has been investigated by Deming [5] with a rather elaborate model. It appears possible to simplify this formulation, although the results may be applicable more to telephoning than to door-to-door interviewing which was Deming's original interest. Consider the problem of estimating a population proportion, call it P. There are PN ones (e.g., persons saying "Yes") and QN zeroes (e.g., persons saying "No" or "Don't know") in the population, where $Q = 1-P$. Suppose that upon being called, a zero has a chance of, say, α of not responding, while the chance of nonresponse for a one is β. Thus the chance of a zero holding out for r calls is α^r, while β^r is the chance that a one will persist as a nonrespondent for r calls.

The sample proportion ones after r calls to a sample of size n, written p_r, is the ratio of two random variables and its expectation, $E(p_r)$, is determined approximately as:

$$(1-Q\alpha^r-P\beta^r)E(p_t)\doteq Q(1-\alpha^r)\times 0+P(1-\beta^r)\times 1=P-P\beta^r \qquad (5.1)$$

As r goes large this tends to P, but short of $r = \infty$ there is a bias in p_r of:

$$PQ(\alpha^r-\beta^r)(1-Q\alpha^r-P\beta^r)^{-1} \quad . \qquad (5.2)$$

Notice, as Deming [5] pointed out, that after r calls the data can be recorded as frequencies and taken as having a multinomial distribution.[2] The model equation for these frequencies may be written as:

$$E(n_{ij}/n) = Q^{1-i}(\alpha^{j-1}-\alpha^j)^{1-i}P^i(\beta^{j-1}-\beta^j)^i \qquad (5.3)$$

where n_{ij} is the number of cases that answer $i=0$ or $i=1$ at $j^{\underline{th}}$ call ($j=1, 2, \ldots, r$). The data can be exhibited as in Table 1, which also shows the residual frequency of nonresponders as n_{r+} along with its model proportion. Ignore γ in Table 1 for now, it will be treated shortly.

Under the supposition that the actually observed frequencies follow a multinomial distribution it becomes simple enough by the method of

scoring (see C. R. Rao [16, p. 165]) to find maximum likelihood estimates of α, β and P . An approximate expression of the variance of the estimate of P when based on r calls can be found by evaluating $\{-E(\partial^2 \ell og\ L/(\partial P)^2)\}^{-1}$ where L is the likelihood of the multinomial distribution. This turns out to give:

$$V(\hat{P}) \doteq n \left(\frac{1-\alpha^r}{Q} + \frac{1-\beta^r}{P} + \frac{(\beta^r-\alpha^r)^2}{Q\alpha^r+P\beta^r} \right)^{-1} , \quad (5.4)$$

which looks plausible since as r goes large the expression tends to

$$n \left(\frac{1}{Q} + \frac{1}{P} \right)^{-1} = PQ/n \quad . \quad (5.5)$$

A cost function that has been used for such surveys charges an amount c_0 for each call and each callback. Then when data are obtained on a case the added cost is c_1 for processing. The expected total survey cost for r calls then becomes:

$$TC=n\{c_0(Q\frac{1-\alpha^r}{1-\alpha} + P\frac{1-\beta^r}{1-\beta})+c_1[Q(1-\alpha^r)+P(1-\beta^r)]\}. \quad (5.6)$$

When, for example, P = .5, α = .9, β = .7, c_0 = \$.50, c_1 = \$1.00 and TC = \$1000 the survey using r = 3 calls and an initial sample of size n = 653 has an estimate with smaller variance than for any other number of calls. If the nonresponse rates are thought to be α = .1 and β = .3 with the other conditions staying the same, then to make one call is optimum according to (5.4). However, in this case no estimate is possible, since there is no way to estimate α separately from β, and so two calls must be made. This would be done by using an initial sample of size n = 645.

A computer program was written to calculate the estimates and their standard errors.[3] One needs only to differentiate the theoretical proportions in Fig. 1 and follow the procedure as shown in Rao [16]. With this approach one makes a test of fit of the model using a chi-square distributed test statistic, X^2 say, on (2r-3) degrees of freedom, where:

$$X^2 = \Sigma(O - E)^2/E \quad , \quad (5.7)$$

where O are observed frequencies in the 2r+1 cells and E are the corresponding theoretical frequencies.

One should not be discouraged by a lack of fit at this stage, since we would expect rather often to find n_{r+} larger than its theoretical frequency due to the presence of never-answer cases. For example, in telephone surveys the additional parameter γ , shown in brackets as optional in Table 1, could represent the proportion of non-working telephone numbers as well as hard-core nonrespondents. In practice one would reset the observed value of n_{r+} , by reducing it, equal to its corresponding theoretical frequency and rerun the fitting routine. This resetting of n_{r+} can be iterated until equality of observed and theoretical frequencies is attained. The difference between the original and the finally fitted proportion in the residual cell is then taken as an estimate of the combined proportion of nonworking numbers and hard-core nonrespondents. Notice that $\hat{P}$ is now an estimate of a conditional proportion, namely the proportion of ones after excluding, for example, the hard-core nonrespondents and non-working numbers.

As an example of the estimation and fitting procedure some data from Kish's textbook [12, p. 544] were examined. These observations resulted from an enumerative survey of gardens in an initial sample of n = 1415 households where r = 3 callbacks were used. The resulting theoretical frequencies are shown, as well as what can be taken as the original counts, in Table 2. The fit statistic, X^2 of (5.7), was .05 on 2 degrees of freedom which indicates that the model can not be faulted.

The parameter estimates were $\hat{P}$ = .442, $\hat{\alpha}$ = .260, $\hat{\beta}$ = .183 and $\hat{\gamma}$ = .158 , and an estimated standard error for $\hat{P}$ was found as .0144. Notice that the estimate of 44% with garden applies now to the sub-population defined by deleting from the frame those households which, no matter how many calls, would never furnish data. These results reinforce the common sense conclusion that the proportion of .447 = 526/1176 having gardens among respondents, fairly well summarizes the data.

6. Optimum Division of Effort Between Increasing Sample Size and Reducing Nonresponse

This brings us to Style 3 in which a preset level of effort to attain response is decided on beforehand and applied to all n cases in the sample. Its drawback is knowledge of a continuum or sequence of methods of steadily increasing efficacy and expense for reducing nonresponse.[4] In theory one can visualize the kind of cost function or plausible relationship between the targeted proportion of nonresponse, to be denoted W_2, and C the cost per case required to be spent in attaining this level of W_2 . It is

$$C = \beta W_2^{-\alpha} \quad , \quad (6.1)$$

and I would judge that β = 1/4 and α = 2 might be reasonable. With these values of β and α a nonresponse proportion of 1/2 would result from spending \$1 per case, while 10% nonresponse would be achieved by spending \$25 and 5% with \$100 . Such a function could only be expected to be realistic for a limited domain and this has perhaps been covered by the numerical values of the example.

An illustration of the use of Style 3 is provided by a sample survey of costs of milk production on dairy farms now going on in North Carolina. A feature of special interest there, and of widespread concern in connection with nonresponse, is the adversary nature of this survey. There are milk producing interests that wish to show how high is the cost of producing milk so as to justify a high price and there are milk consuming interests who wish to demonstrate low costs so that the Milk Commission will lower the price. In such a situation any nonresponse tends to be assigned extreme values, one set for one party and another for the other.

If the variable of interest takes the values of zero or one, or is otherwise limited, then the extremes are straight forward to provide (see

Deming [6, p. 68] and Cochran [4, p. 357]). However, if the variable is, as in this case, numerical and rather open-ended, the following scheme for obtaining the extreme estimates may be considered. In the presence of, say, 20% nonresponse one party would suggest that those 20% were the smallest and so could be balanced by deleting the largest 20% of the observed values and the average then taken. The other party would say that the lowest 20% observed should be deleted. These two estimates based on oppositely censored samples provided a range of uncertainty due to the nonresponse.

A reasonable mediation of these conflicting views might specify that this range be added to the width of a sampling 95 percent confidence interval to furnish a criterion distance to be minimized for fixed total cost by judicious choice of a division of effort between reducing nonresponse and other expenses of increasing sample size. There is an indeterminacy over how many sampling standard errors to combine with the range of nonresponse uncertainty. To be consistent with deleting extremes one should use at least "two sigma" limits and we also show the "three sigma" limits in Table 3. Such a formulation appears close enough to that offered in an article by Birnbaum and Sirken [2] to justify using their symbols: $U = S + b$ where U is total error, S is the familiar 1.96-times-the-sampling-standard-error and b is bias or half that distance between extreme estimates described above.

In connection with the dairy farm survey it is fairly realistic to assume a \$200 data-processing cost per farm and to take $\beta = 1/4$ with $\alpha = 2$, along with a total survey cost of \$15,000. The average reported cost of producing 100 lbs. of milk is around \$10 with a farm-to-farm standard deviation of \$2 . In order to foresee the size of b for varying amounts of percent nonresponse we reason as follows. With, for example, 100 normally distributed observations in hand the smallest is on the average, from sample to sample, 2.50759 standard deviations below the mean. This and other expected values of the normal order statistics (or "normal socres") are taken from tables in Harter [10]. Deleting this smallest value would move the mean upwards by 2.50759/99 = .02533 times the standard deviation. Deleting the two smallest would shift the mean upward by (2.50759 + 2.14814)/98 = .04751 σ's where 2.14814 is the average value of the second largest standard normal observation in 100 . The bias or uncertainty introduced would be (\$2)(.02533) = \$.051 for 1 percent nonresponse and (\$2)(.04751) = \$.095 for 2 percent nonresponse. The calculations in Table 3 use normal scores based on the actual sample sizes and also take account of the binomial variation in number of nonrespondents by finding the expected value of nonresponse uncertainty. Table 3 shows that under such conditions one should aim for W_2 of .05 or .06 and should spend about \$70 to \$100 per case in attaining response. The optimum is quite flat as might be expected.

The survey of milk production costs has been in operation for three years and nonresponse appears to be around 34% this year, even though sample size has been reduced to 50 dairy farms. It remains to be seen whether such procedures as publicizing the study through milk producers associations and having recognized sympathetic authorities to explain how the case of the dairy farmers will be hurt by nonresponse will raise response rates. There is some suspicion that the refusals may be of the, so called, hard core nonresponse type which invalidates the cost function $\beta W_2^{-\alpha}$ and thus renders academic the optimum solution.

One other point of importance in using the results in Table 2 for planning a survey of milk production costs is the perhaps misleading size of Total Uncertainty. One might be distressed that after spending \$15,000 the uncertainty U is still as high as 70¢ or 80¢, which is about 8% of the estimated cost. Part of this is due to the use of two and the even more liberal three sigma ranges which protect against relatively unusual selections, plus the extreme assignment of the nonresponses. A statement of survey precision more consonant with a sample coefficient of variation would be based on further dividing Total Uncertainty, when based on a two sigma range, in half.

7. Further Developments

The pair of direct approaches that were treated in some detail in Sections 5 and 6 can be viewed as special cases of a corresponding pair of more generalized strategies for dealing with survey nonresponse. There are any number of probabilistic models, in addition to the one offered in Section 5, that can be devised to reflect uncertainties about the appearance of cases of nonresponse. Such models need to be worked out and matched against the data. This is nothing more or less than "doing statistics." It is unfortunately true that one thereby looses his grasp of the finite population that is such a conforting concept when the only uncertainties arise from random number tables. However, measurement errors are often so prominent as to demand special attention anyway.

The other generalized strategy that includes the case of optimizing the level of effort to reduce nonresponse may be called the Institutionalization of Surveys Movement. The exercise of certain professions, the legal or the medical say, has become more or less institutionalized within society. Certainly this is true to some extent of census taking, of the conduct of the Agricultural Enumerative Surveys and of the Current Population Survey, as well as of the major public opinion polls. By institutionalization is meant the acceptance of the legitimacy of survey practices within the internalized norms of members of the society.

It is an acceptance that would be built up over a long period of the life of the society and is based on the very tangible advantages of surveys. In order for it to take place it would seem essential that almost all surveys have goals that are clearly seen to be of benefit to the whole society and that they be so carefully designed as to attain their objectives most efficiently. The difficulty in the way of a broad acceptance of surveys is the appearance from time to time of surveys with narrow or confused aims and designed so poorly as to tax a respondent's patience. If such surveys can be made more rare then it may happen that survey interviewing would become a

completely legitimate and morally compelling practice with responding to a survey interview a deeply institutionalized societal norm. Of course, no such idyllic state is in sight but it is a worthwhile goal to pursue since some improvement or even a slowing down of the deterioration in nonresponse rates will be welcome.

FOOTNOTES

[1] For Deming [6] who uses the technique and the word "blank", such selections of non-members of the population are usually done for clerical convenience but the principle is the same.

[2] With such features as stratification and clustering the effective sample size may differ from the number initially selected and thus corrections will need to be made to the standard error calculations.

[3] Copies are available upon request to the author.

[4] That such a continuum may differ from one survey to the next or year to year is to be expected and perhaps one cannot be found. For some rather disappointing results in this direction see Koo et. al. [14].

REFERENCES

[1] Bartholomew, D.J., "A Method of Allowing for Not-at-Home Bias in Sample Surveys," Applied Statistics, 10 (March, 1961),52-59.

[2] Birnbaum, Z.W. and Sirken, M.G., "Bias Due to Non-availability in Sampling Surveys," Journal of the American Statistical Association, 45 (March, 1950), 98-111.

[3] Chapman, David W., "A Survey of Nonresponse Imputation Procedures," Proceedings of the Social Statistics Section, American Statistical Association (1976), Part II, 491-4.

[4] Cochran, William G., Sampling Techniques, Second Ed., New York, John Wiley & Sons, Inc., 1963.

[5] Deming, W. Edwards, "On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Nonresponse and the Bias of Nonresponse," Journal of the American Statistical Association, 48 (December, 1953), 743-772.

[6] Deming, W. Edwards, Sample Design in Business Research, New York, John Wiley & Sons, Inc., 1960.

[7] Ericson, W.A., "Optimal Sample Design with Nonresponse," Journal of the American Statistical Association, 62 (March, 1967), 63-78.

[8] Hansen, M.H. and Hurwitz, W.N., "The Problem of Nonresponse in Sample Surveys," Journal of the American Statistical Association, 41 (December, 1946), 517-529.

[9] Hansen, Morris H., Hurwitz, William N., and Madow, William G., Sample Survey Methods and Theory, Volume 1., New York, John Wiley & Sons, Inc., 1953.

[10] Harter, H. Leon, "Expected Values of Normal Order Statistics," ARL Technical Report 60-292, Aeronautical Research Laboratories, Wright-Patterson Air Force Base, Ohio, (July, 1960).

[11] Kendall, M.G., and Stuart, A., The Advanced Theory of Statistics, Volume 2., Third Ed., New York, Hafner Publishing Co., 1951.

[12] Kish, Leslie, Survey Sampling, New York, John Wiley & Sons, Inc., 1965.

[13] Kish, L. and Hess, I., "A 'Replacement' Procedure for Reducing the Bias of Nonresponse," American Statistician, 13, 4 (October, 1959), 17-19.

[14] Koo, H.P., et al., "An Experiment on Improving Response Rates and Reducing Callbacks in Household Surveys," Proceedings of the Social Statistics Section, American Statistical Association (1976), Part II, 491-4.

[15] Politz, A.N., and Simmons, W.R., "An Attempt to Get the 'Not-at-Homes' into the Sample Without Callbacks," Journal of the American Statistical Association, 44 (March, 1949), 9-31.

[16] Rao, C.R., Advanced Statistical Methods in Biometric Research, New York, John Wiley & Sons, Inc., 1952.

[17] Rubin, Donald B., "Inference and Missing Data," Biometrika, 63, (December, 1976), 581-592.

[18] Sukhatme, P.V., and Sukhatme, B.V., Sampling Theory of Surveys with Applications, Second, Rev. Ed., Ames, Iowa, Iowa State University Press, 1970.

TABLE 1

Model Notation for Observed Frequencies and Theoretical Proportions

Responds at Call No.	Observed Frequencies		Theoretical Proportions	
	Zero	One	Zero	One
1	n_{01}	n_{11}	$Q(1 - \alpha)$	$P(1 - \beta)$
2	n_{02}	n_{12}	$Q(\alpha - \alpha^2)$	$P(\beta - \beta^2)$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
r	n_{0r}	n_{1r}	$Q(\alpha^{r-1} - \alpha^r)$	$P(\beta^{r-1} - \beta^r)$
Residual	n_{r+}		$Q\alpha^r + P\beta^r(+ \gamma)$	
Totals	n		1	

TABLE 2

Original Data and Fitted Frequencies for Responses to a Question on Having a Garden by Number of Visits to the Household Required to Obtain the Response

No. of Call (r =)	Observed Frequencies		Fitted Frequencies	
	No Garden	Had Garden	No Garden	Had Garden
1	489	432	488.42	431.39
2	129	80	128.42	79.40
3	32	14	33.76	14.61
Residual	239		15.3(223.7)	
Totals	1415		1415	

TABLE 3

Total Uncertainty as a Function of Targeted Proportion Nonresponse for Two Levels of Processing Cost and for Two Sigmas and Three Sigmas of Sampling Uncertainty.

Targeted Percent Nonresponse W_2	Expenditure Per Case on Attaining Response $(2\ W_2)^{-2}$	Sample Size, n[a]	Two Sigma Sampling Uncertainty S[a]	Expected Nonresponse Uncertainty b	Total Uncertainty, U = S + b When Processing Cost Per Case is: $200 2σ's	$200 3σ's	$100 2σ's	$100 3σ'a
.01	$ 2500	6	1.633	.030	1.663	2.480	1.663	2.480
.02	625	18	.943	.075	1.018	1.490	.950	1.387
.03	278	31	.718	.120	.839	1.198	.756	1.072
.04	156	42	.617	.163	.780	1.089	.688	.949
.05	100	50	.566	.202	.768	1.051	.669	.900
.06	69	56	.535	.240	.744	1.041	.668	.880
.07	51	60	.516	.276	.792	1.050	.686	.887
.08	39	63	.504	.311	.815	1.067	.703	1.180

[a] Based on a processing cost of $200 per case.

AN INDICATION OF THE EFFECT OF NONINTERVIEW ADJUSTMENT AND POST-STRATIFICATION ON ESTIMATES FROM A SAMPLE SURVEY

Martha J. Banks, University of Chicago

This paper is based on research done at the Center for Health Administration Studies, University of Chicago. Using data from a national sample survey of medical care use in 1970, we investigated various components of total survey error and methods to improve the validity and reliability of the survey estimates. The results of this study will appear in the upcoming book, Total Survey Error: Bias and Random Error in Health Survey Estimates, edited by Ronald Andersen, Judith Kasper, and Martin Frankel. The data were collected and processed by the National Opinion Research Center. The funding for this methodological investigation, as well as for the data collection and basic analysis, was provided by the National Center for Health Services Research. The National Center for Health Statistics also provided valuable support.

This paper concentrates on two features of this investigation, on adjustments for nonresponse and on post-stratification adjustment. Both are relatively easy to implement and so could be used in situations in which other data adjustment techniques might not be felt to be worthwhile.

The basic rationale for nonresponse adjustment might be described as follows: In almost any survey there will be cases which were designated for interview but which were not actually interviewed. Some potential respondents may have refused to be interviewed; others were not at home when repeated interview attempts were made. Making no adjustment for nonresponse implicity assumes that nonrespondents do not tend to differ from respondents in any characteristic of interest. The degree to which they do differ is proportional to the amount of bias introduced by ignoring the nonresponse problem.

Any approach to nonresponse adjustment consists of two elements. First, the population must be categorized into subgroups and the response rate for each group must be determined. The categories chosen to form the subgroups should not only be correlated with characteristics of interest in the study, but be able to be determined without having to obtain the information from the potential respondents themselves.

The second major element in nonresponse adjustment is that of determining the values to impute to the nonrespondents. It usually is reasonable to assume that respondents and nonrespondents falling into the same category tend to have the same characteristics. Sometimes however, we have evidence that respondents and nonrespondents in the same category have measurably different characteristics. This evidence may come from the current study or from external sources. The external data might be from a previous study which either had subsampled nonrespondents or had access to the administrative records of both survey respondents and nonrespondents.

With the data we had available, we chose to try two alternative nonresponse adjustment procedures. Both assume that respondents and nonrespondents within the same category tend to have the same characteristics. Thus the overall results of the two methods differ only because of different choices of categories. The first method creates categories based on geographic location. The second uses information about the reason that no interview was obtained, whether refusal or never-at-home.

Before discussing each method, I first need to discuss the sample used in this investigation. In early 1971, persons in 3880 households were interviewed about the use and cost of health services used during 1970. In all, data were collected for 11,619 individuals. The sample was an area probability sample of the noninstitutionalized population of the continental United States. The sampling procedures oversampled poor persons living in the inner city, persons 65 and over, and rural residents. Naturally, weighting was used to adjust for this oversampling. The weighted nonresponse rate in the survey was 18 percent.

The geographic nonresponse adjustment method used primary sampling unit and sub-sample as the category determinants. This effectively grouped cases within PSUs by the presence of the poor and/or the elderly.

The other method of noninterview adjustment used categories based upon the reason that no interview was completed. To adjust for cases that refused to be interviewed, we increased the weights of those respondents who were not completely cooperative, breaking appointments with the interviewer and so on. The majority of nonrespondents in this study were refusals. To adjust for other types of nonresponses, those due to never being able to find anyone at home, we increased the weights of completed cases according to the number of calls needed to complete the interviews.

Table 1 compares the percentage of interviewed households that received various levels of noninterview adjustment weights according to each method. Most of the sample households were given weights between 1.02 and 2.00 by the geographic method, while most received a weight outside this range from the adjustment based on the reason that no interview was obtained.

TABLE 1 Distribution of sample households by noninterview adjustment factor

INFORMATION USED IN NONINTERVIEW ADJUSTMENT	NONINTERVIEW ADJUSTMENT FACTOR						TOTAL
	1.00	Over 1.00 thru 1.02	Over 1.02 thru 1.15	Over 1.15 thru 1.30	Over 1.30 thru 2.00	Over 2.00 thru 2.40	
Geographic	15.8	0.0	24.8	34.1	24.3	.1	100.0
Reason no interview was completed	67.3	12.7	3.4	1.0	2.9	12.8	100.0

TABLE 2 Effect of nonresponse adjustment on the distribution of sample persons, on estimates of mean number of physician visits for persons seeing a physician, and on estimates of mean hospital expenditure per admission

CHARACTERISTIC	PERCENT OF WEIGHTED SAMPLE PERSONS				MEAN PHYSICIAN VISITS PER PERSON				MEAN EXPENDITURE PER ADMISSION			
		Adjusted for Nonresponse				Adjusted for Nonresponse				Adjusted for Nonresponse		
	Unadjusted	Adjusted Using External Data	Adjusted Using Geographic Information	Adjusted Using Reason No Interview Obtained	Unadjusted	Adjusted Using External Data	Adjusted Using Geographic Information	Adjusted Using Reason No Interview Obtained	Unadjusted	Adjusted Using External Data	Adjusted Using Geographic Information	Adjusted Using Reason No Interview Obtained
Demographic												
Age of oldest family member												
Less than 65 years	86.2%	86.5%	86.2%	86.3%	5.4	5.5	5.4	5.3	$640	$662	$642	$624
65 years or more	13.8	13.4	13.8	13.7	7.8	7.9	7.8	7.7	863	886	877	857
Family income												
Nonpoor	77.0	77.3	77.1	77.8	5.5	5.6	5.6	5.5	674	697	679	643
Poor	23.0	22.8	22.9	22.2	6.5	6.6	6.6	6.5	715	737	717	749
Race												
White	87.9	87.9	88.0	88.3	5.7	5.7	5.7	5.6	684	709	691	659
Nonwhite	12.1	12.1	12.0	11.7	6.0	6.1	5.9	5.9	688	711	669	770
Residence												
Rural nonfarm	24.5	23.7	24.4	24.4	5.4	5.4	5.4	5.3	557	573	565	534
Rural farm	6.8	6.2	6.8	6.6	5.6	5.6	5.6	5.7	575	586	562	574
SMSA central city	29.8	31.2	29.4	29.7	6.0	6.1	6.1	6.0	727	757	745	735
SMSA other urban	26.9	27.0	27.2	26.5	5.6	5.6	5.6	5.4	910	941	904	904
Urban nonSMSA	12.1	12.0	12.2	12.8	6.1	6.2	6.2	6.1	444	458	446	431
Perceived and Evaluated Health												
Perception of health												
Excellent	37.8	-[a]	37.9	38.1	3.6	-[a]	3.6	3.5	488	-[a]	495	475
Good	43.0	-	42.7	42.9	5.3	-	5.3	5.2	609	-	619	614
Fair	12.4	-	12.4	12.1	9.1	-	9.2	8.9	698	-	712	682
Poor	3.9	-	3.9	3.9	14.2	-	14.2	14.5	868	-	877	862
Number of diagnoses												
One	28.2	-	28.2	28.5	3.8	-	3.8	3.8	626	-	643	617
Two	17.6	-	17.7	18.2	5.6	-	5.7	5.8	573	-	578	620
Three	9.2	-	9.1	9.0	8.0	-	8.0	7.8	563	-	559	544
Four or more	9.0	-	9.0	8.4	11.6	-	11.7	11.6	883	-	886	839
Total	100.0%	100.0%	100.0%	100.0%	5.7	5.8	5.8	5.7	$685	$707	$689	$670

[a]Data necessary to provide these estimates are unavailable.

Results from each nonresponse adjustment method were compared with each other and with data unadjusted for nonresponse. These appear in Table 2. Additional columns in this table are labeled "adjusted using external data." The limited number of estimates given in these columns were obtained by using data from various other health surveys to estimated differences between respondents and nonrespondents.

While discussing the data, I would like to stress that the table contains the results we have obtained from using each method of noninterview adjustment. I do not want to make any predictions about what the results would be expected to be if these methods were applied to a number of similar data sets.

Table 2 provides information about the effect of noninterview adjustment on the distribution of sample persons, on the mean number of physician visits for persons with visits, and on the mean hospital expenditure per admission. Most differences in the table are very small. However, all three sections shown the adjustment based on geographic information had less effect on the means than did the adjustment based on the reason that no interview was obtained. The noninterview adjustment seems to have had more effect on hospital expenditures than on physician visits, at least for totals and among the demographic characteristics.

None of this discussion has attempted to suggest which type of adjustment produces the most accurate estimates or even whether or not the time spent doing any type of adjustment for nonresponse is time well spent. In fact in most data collection there is no way to find out what would be the response of all nonrespondents. Therefore there is no way to determine the improvement in the estimates caused by noninterview adjustment. We can only measure the change it makes in the unadjusted estimates.

In our examples few of the estimates adjusted for nonresponse are very different from the unadjusted estimates. However, a well thought-out plan for noninterview adjustment usually is worth making, since doing so is fairly simple. Further, the benefits of noninterview adjustment probably are increasing, since the response rates of most surveys have been declining for at least a decade.

A fairly firm plan for noninterview adjustment should be devised before the study interviews are conducted, so that the desired information used in forming noninterview categories can be collected both for the respondents and for the nonrespondents.

As previously stated, noninterview adjustment also requires values to impute to the nonrespondents in each category. Unless there is firm evidence to the contrary, it would seem best to assume that respondents and nonrespondents falling into the same category are otherwise identical. Doing so usually is preferable to using external data to estimate values to impute to nonrespondents. Definitional and procedural differences between the current data and the external data require caution in adapting the results from the external data sources. It seems unlikely that the cost and time spent locating and adapting external data could often be justified.

It is difficult to choose between the two adjustment methods which use internal data, given the limited information available on the effect of each. I feel that the use of either is somewhat preferable to performing no noninterview adjustment at all, but either set of categories could be used.

Noninterview adjustment is a relatively inexpensive method of reducing nonresponse bias somewhat, but it certainly is no substitute for obtaining actual responses from as many designated respondents as possible. Adjustment should not be used as an excuse for a high nonresponse rate!

The reasons for post-stratification adjustment can be summarized as follows: Compared to the original population, a sample chosen from that population will exhibit chance differences in nearly all possible variables. Usually there are a number of characteristics which are correlated with the dependent variables of interest in the study and for which more reliable estimates exist. Thus the study data generally can be improved by applying a set of factors which adjusts the sample distribution according to the more reliable data.

The more reliable data used for calculating such factors should, of course, be based upon the very same population represented by the sample. Each of the characteristics chosen to form the categories should be fairly highly correlated with statistics of interest.

I have examined the effect of two alternative sets of post-stratification factors. Both were adjusted to Current Population Survey data. The categories used in each appear in Table 3 and in Table 4. The first set adjusts the data according to the CPS distribution of households by race, residence, size, and income. The second set adjusts the data to the distribution of persons by race, sex, and age. I formed the latter set by trying to group sample persons with similar health characteristics. I also considered the weighted and unweighted number of cases per cell. (Nonresponse adjustment should be performed before post-stratification adjustment. Thus the post-stratification factors used with the data presented in Table 2 differ from those given in Table 3.)

Table 5 presents the effect of the use of post-stratification adjustment on our data. This table does not indicate that there was any great change in the data as a result of using either of the sets of post-stratification adjustment factors. Again however, I do not intend to suggest that these specific results would occur if such adjustments were used with any or all similar data.

In order to definitively determine the effect of alternative post-stratification adjustments, we would need the results of a complete census using the sample survey questionnaire. We have had to examine the effect of post-stratification by comparing adjusted and unadjusted estimates from a single sample. Also, we were able to look at only two different types of estimates — that of mean number of physician visits and of mean total hospital expense per admission. (Our attempt to measure the impact of post-stratifica-

TABLE 3 Original post-stratification adjustment categories and factors

CHARACTERISTIC				POST-STRATIFICATION ADJUSTMENT FACTOR	PERCENT OF SAMPLE HOUSEHOLDS	
Race	Residence	Family Size	Household Income		Unweighted, Unadjusted	Weighted, Adjusted
White	SMSA	1	Under $3000	1.622	6.0%	5.9%
White	SMSA	1	$3000 plus	1.191	5.6	8.0
White	SMSA	2+	Under $3000	1.136	3.3	2.5
White	SMSA	2+	$3000-14999	1.160	18.6	29.8
White	SMSA	2+	$15000 plus	1.181	5.6	12.3
White	NonSMSA	1		0.967	5.9	5.7
White	NonSMSA	2+	Under $3000	1.080	3.3	2.7
White	NonSMSA	2+	$3000-14999	0.750	21.9	18.4
White	NonSMSA	2+	$15000 plus	1.147	2.4	3.8
Nonwhite	SMSA	1	Under $3000	1.200	3.3	1.2
Nonwhite	SMSA	1	$3000 plus	0.714	2.2	1.0
Nonwhite	SMSA	2+	Under $3000	0.714	4.3	1.0
Nonwhite	SMSA	2+	$3000-14999	0.600	13.6	4.1
Nonwhite	SMSA	2+	$15000 plus	1.167	0.8	0.7
Nonwhite	NonSMSA	1		1.000	0.6	0.6
Nonwhite	NonSMSA	2+		0.917	2.6%	2.2%

TABLE 4 Alternative post-stratification adjustment categories and factors

CHARACTERISTIC			POST-STRATIFICATION ADJUSTMENT FACTOR	PERCENT OF SAMPLE HOUSEHOLDS	
Race	Sex	Age		Unweighted, Unadjusted	Weighted, Adjusted
White		0 to 5	1.0344	6.5%	8.7%
White		6 to 11	0.9699	7.9	10.2
White		12 to 17	0.9563	8.3	10.2
White	Male	18 to 29	1.1487	5.0	7.8
White	Female	18 to 29	1.1417	5.5	8.3
White	Male	30 to 44	1.0336	5.1	7.3
White	Female	30 to 44	1.0036	5.4	7.5
White		45 to 54	1.1228	6.7	10.3
White		55 to 64	1.0330	6.6	8.3
White		65 to 74	1.0245	6.5	5.5
White		75 plus	1.1384	4.2	3.5
Nonwhite		0 to 8	0.7906	7.3	2.8
Nonwhite		9 to 17	0.6803	8.4	2.6
Nonwhite	Male	18 to 44	0.9781	3.8	2.0
Nonwhite	Female	18 to 44	0.7465	6.0	2.3
Nonwhite		45 to 64	0.7459	4.7	2.0
Nonwhite		65 plus	0.7410	2.3%	0.8%

tion was further confounded by the fact that noninterview adjustment usually would be performed first; while we have had to consider the effect of each separately. Had we computed estimates using combinations of noninterview and post-stratification adjustment, some combination of the two might have interacted in such a way that such adjustments would have had a bigger effect than either individual adjustment would have suggested.) This information on the effect of post-stratification adjustment, limited though it is, is a useful first step in developing a more thorough investigation into the expected effect of such adjustment on different types of data.

Despite the fact that it is difficult to assess the effect of post-stratification on the data and even more difficult to predict what its use would mean to other surveys, I would suggest that it be done. Post-stratification is an extremely inexpensive procedure and should result in at least some small improvement in the data. The choice of categories depends upon the nature of the survey, since the categories should be delineated by characteristics correlated with important estimates in the study.

TABLE 5 Effect of post-stratification adjustment on distribution of sample persons, on estimates of mean number of physician visits for persons seeing a physician, and on estimates of mean hospital expenditure per admission

	PERCENT OF WEIGHTED SAMPLE PERSONS			MEAN PHYSICIAN VISITS PER PERSON			MEAN EXPENDITURE PER ADMISSION		
	Without Post-Stratification Adjustment	With Post-Stratification Adjustment		Without Post-Stratification Adjustment	With Post-Stratification Adjustment		Without Post-Stratification Adjustment	With Post-Stratification Adjustment	
		Original Categories	Alternative Categories		Original Categories	Alternative Categories		Original Categories	Alternative Categories
Demographic									
Age of oldest family member									
Less than 65 years	86.8%	86.2%	86.5%	5.4	5.4	5.4	$616	$640	$616
65 years or more	13.3	13.8	13.5	7.7	7.8	7.7	830	863	831
Family income									
Nonpoor	75.7	77.0	77.1	5.5	5.5	5.6	649	674	647
Poor	24.3	23.0	22.9	6.3	6.5	6.3	685	715	691
Race									
White	83.8	87.9	87.5	5.7	5.7	5.7	652	684	650
Nonwhite	16.2	12.1	12.5	5.9	6.0	5.9	702	688	730
Residence									
Rural nonfarm	25.8	24.5	26.5	5.4	5.4	5.4	544	557	540
Rural farm	7.8	6.8	7.9	5.5	5.6	5.5	555	575	566
SMSA central city	29.1	29.8	27.7	6.0	6.0	6.0	706	727	711
SMSA other urban	23.3	26.9	23.8	5.5	5.6	5.5	920	910	915
Urban nonSMSA	13.9	12.1	14.2	6.1	6.1	6.2	447	444	447
Perceived and Evaluated Health									
Perception of health									
Excellent	37.0	37.8	37.4	3.5	3.6	3.6	467	488	472
Good	43.3	43.0	43.1	5.2	5.3	5.3	600	609	596
Fair	12.7	12.4	12.7	8.9	9.1	9.0	670	698	682
Poor	3.8	3.9	3.8	14.1	14.2	14.2	816	868	809
Number of diagnoses									
One	27.9	28.2	28.0	3.8	3.8	3.8	605	626	600
Two	17.5	17.6	17.6	5.6	5.6	5.7	573	573	574
Three	8.9	9.2	9.2	8.1	8.0	8.0	543	563	540
Fouror more	8.6	9.0	8.9	11.6	11.6	11.5	836	883	839
Total	100.0%	100.0%	100.0%	5.7	5.7	5.7	$658	$685	$658

ON THE REDUCTION OF RESPONSE BIAS IN SAMPLE SURVEYS USING AUXILIARY INFORMATION

R. P. Chakrabarty, Jackson State University

SUMMARY

The control and reduction of response biases is often a major problem in sample surveys. In this paper, we develop a method for reducing response biases by using auxiliary information. When an auxiliary variable 'x' that is correlated with the variable of interest 'y' is available it is shown that the classical ratio estimator of the population mean or total of y has less response bias than the estimator that uses y - information only.

The ratio estimator, however, does not help much when the response bias for y and/or x is very large. In such situations the use of a double sampling method is useful. For a random sub-sample of the original sample either true values of y, x or values that have less biases than in the original samples are obtained. A difference estimator computed from two samples is shown to be very effective in reducing response biases.

Key Words: Response bias, ratio estimator, double sampling

1. INTRODUCTION

The general theory of sample surveys assumes that the observation y_i on the ith unit in the sample is the "true value" for that unit. The variance of an estimate obtained from the sample is assumed to arise solely from the random sampling variation that is present when only n units in the sample are measured out of the N in the population. By implication, we assume that in the case of a census (n=N) we obtain the "true value" of the mean or total of the population. In practice, however, in most surveys different types of "non-sampling errors" such as "non-response" (failure to measure some of the units in the sample), measurement error or "response error" (respondents giving in-accurate information) may be present. In this paper we are not concerned with the problem of non-response. For literature on non-response see Cochran (1977).

"Response error", in the broadest sense, means the errors that might arise from faulty measurements and observations, in-accurate answers by respondents and "interviewer bias" etc. First, we outline briefly the general theory of response errors from Madow (1965).

Let y_i be the "true value" of the character y for the ith unit of the population and y_i^* be the random variable that is the choice of the respondent i, if the respondent i is asked the question for which true value for that respondent is y_i. Let $E(y_i^*) = a_i$. Then, for the ith respondent, the response bias and the variance of the response are

$$B_i = a_i - y_i$$

and

$$V_i = E(y_i^* - a_i)^2$$

respectively. Finally, MSE of response is

$$M_i = E(y_i^* - y_i)^2 = E(y_i^* - a_i)^2 + (a_i - y_i)^2.$$

The objective of the survey is to estimate the "true population mean" $\bar{Y}$, of y and a sample of size n is drawn from the N units in the population.

It can be shown that MSE of $\bar{y}^*$ (ignoring the fpc) is

$$MSE(\bar{y}^*) = \frac{S_a^2}{n} + \frac{\sigma_w^2}{n}\left[1 + (n-1)\rho_w\right] + (\bar{A}-\bar{Y})^2 \qquad (1.1)$$

where

$$\sigma_w^2 = \sum_1^N \sigma_i^2/N, \quad \sigma_i^2 = \sigma_{y_i^*}^2,$$

$$\rho_w \sigma_w^2 = \frac{1}{N(N-1)} \sum_{i \neq j} E(y_i^* - a_i)(y_j^* - a_j)$$

and

$$\bar{A} = \sum_1^N a_i/N.$$

The formula (1.1) contains two terms S_a^2/n and $\sigma_w^2(1-\rho_w)/n$ that decrease as n increases. The remaining two terms $\sigma_w^2 \rho_w$ and $(\bar{A}-\bar{Y})^2$ are independent of n. Thus in large samples the MSE is likely to be dominated by these two terms, the ordinary sampling variance becoming unimportant and misleading as a guide to the real accuracy of the results. These results emphasize the importance of discovering and controlling response errors in sample surveys.

In recent years much of the research on sampling practice has been devoted to the study of response errors. Cochran (1977) has given an extensive discussion of this topic. Madow (1965) has suggested the use of double sampling technique using y information only for eliminating or reducing the response bias. In this paper, we develop a method for reducing response biases by using auxiliary information.

2. General Statement of the Problem

Consider a finite population of N units. Let y_i and x_i denote the true values of characteristic of interest y and auxiliary characteristic x respectively attached to the i th unit of the population. The parameter to be estimated is the population mean of y, $\bar{Y} = \sum_{i=1}^N y_i/N$. or population to total $Y = N\bar{Y}$.

From a simple random sample of $n(\leq N)$ units we have the sample data (y_i^*, x_i^*), i = 1,2,...n. Note that y_i^* and x_i^* are the values reported by respondents instead of true values (y_i, x_i). The estimator of Y is

$$\hat{Y} = N\bar{y}^* \qquad (2.1)$$

that uses y information only. The ratio estimator of Y is

$$\hat{Y}_r = N \bar{y}_r^* \qquad (2.2)$$

where $\bar{y}_r^* = (\bar{y}^* / \bar{x}^*)\, \bar{X}$

and $\bar{y}^*$ and $\bar{x}^*$ are sample means of y^* x^* respectively.

We note that both estimators $\hat{Y}$ and $\hat{y}_r$ will have bias due response errors. The ratio estimators will also have the usual bias of a ratio estimator that occurs because only a fraction of the population is sampled. We shall ignore the usual bias by assuming sample size n is sufficiently large and investigate the bias due to response errors. An interesting case of response errors over-reporting was found to have occured in Agricultural Surveys in Texas. To fix the idea, let us consider the agricultural surveys in Texas. After the A. S. C. S. list in each county has been consolidated, a random sample of n operator addresses will be drawn out of N addresses and data will be collected by mail questionnaire. For simplicity assume that there is 100% response to mail questionnaire. If over-reporting occurs both y_i^* and x_i^* refer to an operation which is in excess of that properly due to the i th operator of the A.S.C.S. list. Such over-reporting for certain units in the sample might be caused by the fact that the "frame" (list of units) is out of date. Over reporting can sometimes be detected by scrutiny of data and corrected by direct interview. Such a procedure is rather costly and, therefore, seldom feasible in large scale surveys. Hartley (1966) proposed the use of the ratio estimator (2.2) to eliminate the bias due to over-reporting in agricultural surveys in Texas. In this paper, we show that the ratio estimator has less response bias than the estimator that uses y-information only. The ratio estimator suggested by Hartley, however, does not help much when the response bias for y and/or x is very large. In such situations a difference estimator obtained by a double sampling method is shown to be very effective in reducing response biases.

3. Over-reporting Bias Under a Model

Conceptually, we can imagine that a large number of independent repetitions of the measurement on each unit of the population are possible. Let $y_{i\alpha}^*$ and $x_{i\alpha}^*$ be the values of the characters y and x obtained for the ith unit in the αth repetition. Then we have the model

$$y_{i\alpha}^* = y_i + \eta_{i\alpha} \qquad (3.1)$$

$$x_{i\alpha}^* = x_i + \xi_{i\alpha}$$

where, as before, y_i and x_i denote the true values of the characters y and x for the ith unit and $\eta_{i\alpha}$ and $\xi_{i\alpha}$ are errors of reporting in the αth repetition. If there is no over-reporting for the ith unit in the α th repetition then $\eta_{i\alpha} = \xi_{i\alpha} = 0$; otherwise $\eta_{i\alpha} > 0$ and $\xi_{i\alpha} > 0$. Under the repeated measurement of the units we have

$$\varepsilon\{\eta_{i\alpha} \mid y_i\} = \mu_{1i};\ \varepsilon\{\xi_{i\alpha} \mid x_i\} = \mu_{2i}$$

where ε denotes the expectation over repeated measurements.

Over-reporting bias is essentially a non-sampling error in the sense that the bias is not eliminated even in the census. In this section we, therefore, confine ourselves to the bias of the estimator $\bar{Y}_r^* = (\bar{Y}^* / \bar{X}^*)\, \bar{X}$, (rather than that of $\bar{y}_r^*$ under sampling), under the above model, to simplify the discussion.

$$\text{Now } \bar{Y}^*/\bar{X}^* = \sum_{i=1}^{N} (y_i + \eta_{i\alpha}) / \sum_{i=1}^{N} (x_i + \xi_{i\alpha})$$

$$= \frac{\bar{Y}}{\bar{X}} \left(1 + \frac{\bar{\eta}_\alpha}{\bar{Y}}\right) \left(1 + \frac{\bar{\xi}_\alpha}{\bar{X}}\right)^{-1} \qquad (3.2)$$

where $\bar{\eta}_\alpha = N^{-1} \sum_{i=1}^{N} \eta_{i\alpha}$ and $\bar{\xi}_\alpha = N^{-1} \sum_{i=1}^{N} \xi_{i\alpha}$.

Assuming $\bar{\xi}/\bar{X} < 1$, which will be generally true, and using Taylor's expansion for $(1 + \bar{\xi}_\alpha / \bar{X})^{-1}$ we get, neglecting cubic and higher order terms in $\bar{\xi}/\bar{X}$

$$\frac{\bar{Y}^*}{\bar{X}^*} \doteq \frac{\bar{Y}}{\bar{X}} \left\{1 + \frac{\bar{\eta}_\alpha}{\bar{Y}} - \frac{\bar{\xi}_\alpha}{\bar{X}} - \frac{\bar{\eta}_\alpha \bar{\xi}_\alpha}{\bar{Y}\,\bar{X}} + \frac{\bar{\xi}_\alpha^2}{\bar{X}^2}\right\}. \qquad (3.3)$$

Further,

$$\varepsilon(\bar{\eta}_\alpha/\bar{Y}) = \bar{\mu}_1/\bar{Y},\ \varepsilon(\bar{\xi}_\alpha/\bar{X}) = \bar{\mu}_2/\bar{X},$$

$$\varepsilon(\bar{\eta}_\alpha \bar{\xi}_\alpha/\bar{Y}\bar{X}) = \{\mathrm{Cov}(\bar{\eta}_\alpha, \bar{\xi}_\alpha) + \mu_1\mu_2\}/\bar{Y}\,\bar{X}$$

$$= \mu_1\,\mu_2\,(1 + \rho^*\, C_1\, C_2)/\,\bar{Y}\,\bar{X},$$

$$\varepsilon(\bar{\xi}_\alpha^2/\bar{X}^2) = \bar{\mu}_2^2\,(1 + C_2^2)/\,\bar{X}^2,$$

where $\bar{\mu}_1 = N^{-1} \sum_{i=1}^{N} \mu_{1i}$, $\bar{\mu}_2 = N^{-1} \sum_{i=1}^{N} \mu_{2i}$, $C_1^2 = V(\bar{\eta}_\alpha)/\bar{\mu}_1^2$,

$C_2^2 = V(\bar{\xi}_\alpha)/\bar{\mu}_2^2$ and ρ^* is the coefficient of correlation between $\bar{\eta}_\alpha$ and $\bar{\xi}_\alpha$.

Thus

$$\varepsilon\left(\frac{\bar{Y}^*}{\bar{X}^*}\right) \doteq \frac{\bar{Y}}{\bar{X}} \left\{1 + \frac{\bar{\mu}_1}{\bar{Y}} - \frac{\bar{\mu}_2}{\bar{X}} + \frac{\bar{\mu}_2^2}{\bar{X}^2}(1 + C_2^2) - \frac{\bar{\mu}_1\bar{\mu}_2}{\bar{Y}\,\bar{X}}(1 + \rho^*\, C_1 C_2)\right\}. \qquad (3.4)$$

The relative bias of $\bar{y}_r^*$ as an estimator of $\bar{Y}$ is

$$B=\bar{X}\{\varepsilon(\bar{y}^*/\bar{x}^*)-\bar{Y}/\bar{X}\}/\bar{Y}\doteq \frac{\bar{\mu}_1}{\bar{Y}} - \frac{\bar{\mu}_2}{\bar{X}} + \frac{\bar{\mu}_2^2}{\bar{X}^2}(1+C_2^2) - \frac{\bar{\mu}_1\bar{\mu}_2}{\bar{Y}\,\bar{X}}(1+\rho^* C_1 C_2). \quad (3.5)$$

If $\bar{\mu}_1/\bar{Y}=\bar{\mu}_2/\bar{X}$, (3.5) reduces to $B\doteq \bar{\mu}_2^2(C_2^2 -\rho^* C_1 C_2)/\bar{X}^2$.

It is of interest to investigate the magnitude of the relative bias. The relative bias clearly depends on the values of the parameters ρ^*, C_1 C_2, $\bar{\mu}_1/\bar{Y}$ and $\bar{\mu}_2/\bar{X}$. We now make the reasonable assumption

$$C_1=C_2 =C \text{ (say)} \quad (3.6)$$

to simplify the discussion and let

$$\bar{\mu}_2/\bar{X} = L\bar{\mu}_1/\bar{Y} =L\ P \quad \text{(say)} \quad (3.7)$$

where P is the relative bias of the estimator $\bar{y}^*$ which does not use the x-information and L is the ratio of the relative bias of $\bar{x}^*$ to that of $\bar{y}^*$. Then

$$B\doteq(1-L)\ (1-LP)\ P+P^2C^2\ L(L-\rho^*). \quad (3.8)$$

We have made a numerical evaluation of the magnitude of the relative bias for different values of parameters L,P,C and ρ^*. The coefficient of variation C is of order $N^{-1/2}$ if the measurements on different units are uncorrelated; otherwise C could be large. We have, therefore, included small and large values of C to cover both the cases. The results are presented in Tables 1,2 and 3.

The following conslusions may be drawn from the Tables 1, 2 and 3.

(1) For P, the relative bias of the estimator $\bar{y}^*$, not exceeding 25%, $|B|$, the relative bias of the ratio estimator $\bar{y}_r^*$, is less than P. When P=50%, $|B|< P$ if $C \leq 1.50$. These results demonstrate the effectiveness of the ratio estimator in reducing the over-reporting bias.

(2) For fixed $L \leq 1.0$, P and C, $|B|$ decreases as ρ^* increases.

(3) For fixed L, C and ρ^*, $|B|$ increases with P.

(4) The relative bias is, for all practical purposes, negligible ($\leq 5\%$) for $.75 \leq L \leq 1.25$ and $C \leq 2.50$ if $P \leq 10\%$ even when the correlation is low ($\rho^* = .3$).

(5) For $.75 \leq L \leq 1.25$, $|B|$ is less than 6% if $P \leq 25\%$, $C \leq 1.50$ and $\rho^* > .5$.

(6) The relative bias becomes, in general, serious with P >25% and C >1.50. In such cases, it is higher for L >1 than L <1.

4. The Elimination or Reduction of Over-reporting Bias by Double Sampling

In the previous section we have shown that the ratio estimator suggested by Hartley, is generally effective in reducing the over-reporting bias. The ratio estimator, however, does not help much when the relative bias for the character 'y' and/or for the character 'x' is large. In such situations, the use of a double sampling method seems to be appropriate. In this section we, therefore, outline the double sampling technique for our present problem.

A random subsample of size n_1 ($< n$) out of the original sample of size n is drawn. The "true values" (y_i, x_i) are ascertained for the operators selected in the subsample either from records if that is feasible or by interviewing the selected operators. We note that true values may not always be obtained by this method but the values obtained will have smaller biases than the values (y_1^*, x_1^*) reported by the respondents. However, we suppose that the true values (y_i, x_i) are obtained for the subsample to simplify the discussion. We also assume, to simplify the discussion, that y_1^*, and x_1^* are fixed quantities, instead of random variables.

Let $\bar{y}_1^*$, $\bar{x}_1^*$, $\bar{y}_1$ and $\bar{x}_1$ be the subsample means. Then an estimator of $\bar{Y}$ is given by

$$\bar{y}_r' = t\,\bar{X} \quad (4.1)$$

where

$$t=\frac{\bar{y}^*}{\bar{x}^*} - \left(\frac{\bar{y}_1^*}{\bar{x}_1^*} - \frac{\bar{y}_1}{\bar{x}_1}\right) \quad (4.2)$$

is the difference estimator. Clearly the expected value of t is

$$E(t) \doteq R$$

and

$$E(\bar{y}_r') \doteq \bar{Y}$$

provided n_1 is sufficiently large i.e., $\bar{y}_r'$ is approximately unbiased. The variance of t is given by

$$V(t)=V\left(\frac{\bar{y}^*}{\bar{x}^*}\right)+V\left(\frac{\bar{y}_1^*}{\bar{x}_1^*} - \frac{\bar{y}_1}{\bar{x}_1}\right) -2\ \mathrm{Cov}\left(\frac{\bar{y}^*}{\bar{x}^*}, \frac{\bar{y}_1^*}{\bar{x}_1^*} - \frac{\bar{y}_1}{\bar{x}_1}\right). \quad (4.3)$$

Now, the variance of $\bar{y}^*/\bar{x}^*$ is

$$V\left(\frac{\bar{y}^*}{\bar{x}^*}\right) \doteq \frac{1}{n\,\bar{X}^{*2}} S^2(y^* -R^* x^*) \quad (4.4)$$

(neglecting fpc), where

$$R^* = \bar{Y}^*/\bar{X}^*. \quad (4.5)$$

Using conditional expections, it can be shown that

$$V\left(\frac{\bar{y}_1^*}{\bar{x}_1^*} - \frac{\bar{y}_1}{\bar{x}_1}\right) \doteq \frac{n-n_1}{n\,n_1}\,\frac{1}{\bar{X}^{*2}}\,S^2_{(y^*-R^*x^*)}$$

$$+\frac{1}{n\,\bar{X}^{*2}}\,S^2_{(y^*-R^*x^*)}+\frac{n-n_1}{n\,n_1}\,\frac{1}{\bar{X}^2}\,S^2_{(y-Rx)}$$

$$+\frac{1}{n\,\bar{X}^2}\,S^2_{(y-Rx)}-\frac{n-n_1}{n\,n_1}\,\frac{2}{\bar{X}^*\bar{X}}\,S_{(y^*-R^*x^*)(y-Rx)}$$

$$-\frac{2}{n\bar{X}^*\bar{X}}\,S_{(y^*-R^*x^*)(y-Rx)} \qquad (4.6)$$

and

$$\mathrm{Cov}\left(\frac{\bar{y}^*}{\bar{x}^*},\ \frac{\bar{y}_1^*}{\bar{x}_1^*} - \frac{\bar{y}_1}{\bar{x}_1}\right) \doteq \frac{1}{n\bar{X}^{*2}}\,S^2_{(y^*-R^*x^*)}$$

$$-\frac{1}{n\bar{X}^*\bar{X}}\,S_{(y^*-R^*x^*)(y-Rx)} \qquad (4.7)$$

Substitution of (4.4), (4.6), and (4.7) into (4.3) leads to

$$V(t)\doteq \frac{1}{n\bar{X}^2}\,S^2_{(y-Rx)}+\frac{n-n_1}{n\,n_1}\left\{\frac{1}{\bar{X}^{*2}}\,S^2_{(y^*-R^*x^*)}\right.$$

$$\left.+\frac{1}{\bar{X}^2}\,S^2_{(y-Rx)}-\frac{2}{\bar{X}^*\bar{X}}\,S_{(y^*-R^*x^*)(y-Rx)}\right\}$$

$$\doteq \frac{1}{n\bar{X}^2}\,S^2_{(y-Rx)}+\frac{n-n_1}{n\,n_1}\,S^2_{(r^*-r)}\ \text{(say)}. \qquad (4.8)$$

Finally, we obtain the variance of $\bar{y}_r'$ as

$$V(\bar{y}_r')=\bar{X}^2\left\{\frac{1}{n\bar{X}^{*2}}\,S^2_{(y-Rx)}+\frac{n-n_1}{n\,n_1}\,S^2_{(r^*-r)}\right\}. \qquad (4.9)$$

If no subsample is selected, then from the original sample of size n the estimator of $\bar{Y}$ is $\bar{y}_r^*$ given by (4.2). The over-reporting bias of $\bar{y}_r^*$, ignoring the technical bias of the ratio estimator, is given by

$$B^* = \bar{X}(R^*-R) = \bar{Y}^* - \bar{Y} \qquad (4.10)$$

and the MSE of $\bar{y}_r^*$ is

$$MSE(\bar{y}_r^*)\doteq \bar{X}^2\left\{\frac{1}{n\bar{X}^{*2}}\,S^2_{(y^*-R^*x^*)}+(R^*-R)^2\right\}. \qquad (4.11)$$

To compare the double sampling plan with single sampling we assume

$$\frac{1}{n\bar{X}^2}\,S^2_{(y-Rx)} = \frac{1}{n\bar{X}^{*2}}\,S^2_{(y^*-R^*x)},$$

to simplify the discussion. Then from (4.9) and (4.11) we have

$$\Delta=MSE(\bar{y}_r^*)-V(\bar{y}_r')=\bar{X}^2\left\{(R^*-R)^2-\frac{n-n_1}{n\,n_1}\,S^2_{(r^*-r)}\right\}$$

$$=\left\{(\bar{Y}^*-\bar{Y})^2-\frac{n-n_1}{n\,n_1}\,\bar{X}^2\,S^2_{(r^*-r)}\right\} \qquad (4.12)$$

If $B^* = \bar{Y}^*-\bar{Y}$ is small, then Δ is small or negative there is little to gain from the double sampling. As we have stated earlier, the double sampling technique will be used only when the over-reporting bias B^* is large. When B^* is large, then if $S^2_{(r^*-r)}$ is not large compared to $(R^*-R)^2$ it would be possible with a moderate value of n_1 to make

$$\frac{n-n_1}{n\,n_1}\,\bar{X}^2\,S^2_{(r^*-r)} < (0.1)\,(\bar{Y}^*-\bar{Y})^2;$$

even a smaller multiplier than 0.1 should not be difficult to attain. Thus the double sampling technique would be effective in reducing the bias. The costs of obtaining the true values (y_i, x_i) for the subsample may be quite high compared to the costs of collecting data by mail questionnaire for the large sample and still the double sampling technique would be efficient.

Considering appropriate cost functions for data collection by mail questionnaire and interviews and using the variance formulas for single sampling and double sampling, one could formulate an optimum double sampling scheme. We are at present working on this and hope to report the results in a subsequent paper.

REFERENCES

Cochran, W. G. (1977). "Sampling Techniques", John Wiley and Sons, New York.

Hartley, H. O. (1966). "Brief Description of Unbiased A. S. C. S. List Procedures" Unpublished manuscript.

Madow, W. G. (1965). "On Some Aspects of Response Error Measurement", Proceedings of American Statistical Association (Social Statistics Section), 182-92.

Table 1. Over-reporting Bias of the Ratio Estimator for L=.75 and Selected Values of P, C and ρ.

L = 0.75

P%	C	Bias (absolute value %)			
		ρ=.3	ρ=.5	ρ=.7	ρ=.9
10	.10	2.32	2.31	2.31	2.31
	.50	2.40	2.36	2.32	2.28
	1.00	2.65	2.50	2.35	2.20
	1.50	3.07	2.73	2.40	2.06
	2.00	3.66	3.06	2.46	1.86
	2.50	4.42	3.48	2.55	1.61
25	.10	5.10	5.09	5.08	5.07
	.50	5.61	5.37	5.14	4.90
	1.00	7.12	6.25	5.31	4.37
	1.50	9.82	7.71	5.61	3.50
	2.00	13.52	9.77	6.02	2.27
	2.50	18.26	12.40	6.54	.68
50	.10	7.90	7.86	7.82	7.78
	.50	9.92	8.98	8.05	7.11
	1.00	16.25	12.50	8.75	5.00
	1.50	26.80	18.36	9.99	1.48
	2.00	41.56	26.56	11.56	3.44
	2.50	60.55	37.11	13.67	9.77

Table 2. Over-reporting Bias of the Ratio Estimator for L=1.00 and Selected Values of P, C and ρ.

L = 1.00

P%	C	Bias (absolute value %)			
		ρ=.3	ρ=.5	ρ=.7	ρ=.9
10	.10	.007	.005	.003	.001
	.50	.17	.12	.07	.02
	1.00	.70	.50	.30	.10
	1.50	1.58	1.12	.67	.22
	2.00	2.80	2.00	1.20	.40
	2.50	4.38	3.12	1.87	.62
25	.10	.04	.03	.02	.01
	.50	1.09	.78	.47	.16
	1.00	4.37	3.12	1.87	.62
	1.50	9.84	7.03	4.22	1.41
	2.00	17.50	12.50	7.50	2.50
	2.50	27.34	19.53	11.72	3.91
50	.10	.17	.12	.07	.02
	.50	4.37	3.12	1.87	.62
	1.00	17.50	12.50	7.50	2.50
	1.50	39.37	28.12	16.87	5.62
	2.00	70.00	50.00	30.00	10.00
	2.50	109.37	78.12	46.87	15.62

Table 3. Over-reporting Bias of the Ratio Estimator for L=1.25 and Selected Values of P, C and ρ.

L = 1.25

P%	C	Bias (absolute value %)			
		ρ=.3	ρ=.5	ρ=.7	ρ=.9
10	.10	2.18	2.18	2.18	2.18
	.50	1.89	1.95	2.01	2.07
	1.00	1.00	1.25	1.50	1.75
	1.50	.48	.08	.64	1.20
	2.00	2.50	1.56	.56	.43
	2.50	5.23	3.67	2.11	.55
25	.10	4.22	4.24	4.25	4.26
	.50	2.44	2.83	3.22	3.61
	1.00	3.12	1.56	0.00	1.56
	1.50	12.40	8.89	5.37	1.86
	2.00	25.39	19.14	12.89	6.64
	2.50	42.08	32.32	22.56	12.79
50	.10	4.39	4.45	4.52	4.57
	.50	2.73	1.17	.39	1.95
	1.00	25.00	18.75	12.50	6.25
	1.50	62.11	48.05	33.98	19.92
	2.00	114.06	87.06	64.06	39.06
	2.50	180.86	141.80	102.73	63.67

SOME INTERESTING RESULTS IN UNEQUAL PROBABILITY SAMPLING FROM A FINITE POPULATION

Grant Capps, U. S. Bureau of the Census

The purpose of this paper is two-fold. First, several results are presented for dealing with the most general of unequal probability sampling schemes. These results are considerably more general than presented in most texts, which generally only deal with the two special cases of unequal probability with and without replacement sampling schemes. The first stage of selection in the Current Population Survey as conducted by the U. S. Bureau of the Census provides a useful application of the general theory. Second, for the common within stratum sample size of n=2, this paper proposes a simple sample selection method that attempts to serve as a compromise between the two frequently opposing survey requirements of a small true variance and an unbiased and fairly stable estimate of that variance. Essentially, this new sampling scheme makes use of both the a-priori information used in the strata formation and a well-known unequal probability without replacement selection method. For the proposed scheme, two estimators of the population total are considered and compared both theoretically and empirically.

I. GENERALIZED UNEQUAL PROBABILITY SAMPLING FROM A FINITE POPULATION

As is well known, the popular unequal probability with and without replacement sampling schemes are special cases of a much more general sampling scheme. In the following sections, the general theory associated with this general sampling scheme is developed. Please note, it is not claimed that each of the general results about to be presented are necessarily new and original; however, some of these results are at best not very well-known, while others are included for completeness.

A. General Sampling Scheme. Suppose it is required to select a sample for the purpose of estimating some unknown population total. In general, the sampler is free to assign varying probabilities (including zero) to each possible sample configuration. Let there be N population units and suppose we wish to select a sample of size n, not necessarily distinct, units, where n is a fixed constant. The i^{th} population unit has a known variate (or measure of size) x_i and an unknown variate (characteristic of interest) y_i associated with it (i=1=1,2,...,N).

Let $Y = \sum_{1}^{N} y_i$, $X = \sum_{1}^{N} x_i$, and $P_i = x_i/X$ (i=1,2,...,N).

We seek to estimate the unknown population total Y by selecting a sample of size n using some well-defined sampling scheme.

Denote by t_i (i=1,2,...,N) the number of times the i^{th} unit is included in the chosen sample. A technique originally proposed by Cornfield [3], and used by both Cochran [2] and Raj [7] in their excellent sampling tests when handling the special cases of with and without replacement sampling, is to treat the t_i (i=1,...,N) as the random variables rather than the y_i (i=1,2,...,n) where here y_i is the value of the characteristic for the i^{th} unit selected in sample. Raj [7] went slightly further and proposed using the "t_i" technique for any general sample design. However, he did not present the relevant results, as done here. The sampling scheme itself will uniquely determine the joint probability distribution of the t_i. As will be shown, for the quantities usually estimated, the joint distribution of the t_i is not required. All that is generally required, are the two marginal probabilities, $Pr(t_i)$ and $Pr(t_i,t_j)$ for $i \neq j$, which permit the computation of $E(t_i)$ and $E(t_i t_j)$.

Clearly, since n is fixed, we have

$$n = \sum_{1}^{N} t_i = \sum_{1}^{N} E(t_i) \quad \text{and} \tag{1}$$

$$\mathrm{Cov}(t_i,t_i) = \mathrm{Cov}\left(t_i,\ n - \sum_{j\neq i}^{N} t_j\right) = -\sum_{j\neq i}^{N} \mathrm{Cov}(t_i,t_j) \tag{2}$$

The nature of the sampling scheme employed will determine the difficulty involved in computing $E(t_i)$ and $\mathrm{Cov}(t_i,t_j)$ which are assumed to exist.

B. An Unbiased Estimator for Y. The generalized estimator for Y considered here is:

$$\hat{Y} = \sum_{1}^{N} (y_i) \frac{t_i}{E(t_i)} = \sum_{1}^{n} \frac{y_i}{E(t_i)} \ . \tag{3}$$

The usual assumption that $E(t_i)>0$ (i=1,2,...,N) has been implicily made here. If the variable of interest (y_i) and the measures of size (x_i) are highly correlated, then we generally desire $E(t_i)$ to be proprotional to x_i (i=1,2,...,N). $\hat{Y}$ is clearly an unbiased estimator for Y, since

$$E(\hat{Y}) = \sum_{1}^{N} (y_i) \frac{E(t_i)}{E(t_i)} = Y.$$

There are many unbiased estimators, some possibly better and others worse, however, this paper addresses only the above estimator, as it is the general extension of the classical with and without replacement (Horvitz-Thompson) [6] estimators. The variance of $\hat{Y}$ and two unbiased variance estimators will now be derived.

C. The Variance of $\hat{Y}$. The sampling variance of $\hat{Y}$ can be expressed in two different, though algebraically, equivalent ways. The straightforward expression for $V(\hat{Y})$ is the quadratic form given by

$$V(\hat{Y}) = \mathrm{Cov}(\hat{Y},\hat{Y}) = \sum_{i}^{N}\sum_{j}^{N} \frac{y_i}{E(t_i)} \frac{y_j}{E(t_j)} \mathrm{Cov}(t_i,t_j) \ . \tag{4}$$

Using (2) $V(\hat{Y})$ can be alternatively expressed as:

$$V(\hat{Y}) = -\sum_{i<j}^{N}\sum^{N} \mathrm{Cov}(t_i,t_j)[\Delta y_{ij}^2], \tag{5}$$

$$\text{where } \Delta y_{ij} = \left[\frac{y_i}{E(t_i)} - \frac{y_j}{E(t_j)}\right].$$

D. Two Unbiased Variance Estimators. Two different variance estimators are suggested by (4) and (5) whenever $E(t_i t_j)>0$ for all distinct pairs $i \neq j$. From (4), it is clear that an unbiased estimator of the sampling variance V(Y)

is given by

$$v_1(\hat{Y}) = \sum_{i\neq j}^{N}\sum^{N} \frac{t_i t_j}{E(t_i t_j)} \frac{y_i}{E(t_i)}\left[\frac{y_j}{E(t_j)}\right] \mathrm{Cov}(t_i,t_j) + \sum_{i}^{N} \frac{t_i}{E(t_i)}\left[\frac{y_i}{E(t_i)}\right]^2 \mathrm{Cov}(t_i,t_i) \quad (6)$$

From (5) it is obvious that another unbiased estimator for $V(\hat{Y})$ is

$$v_2(\hat{Y}) = -\sum_{i\neq j}^{N}\sum^{N} \frac{t_i t_j}{E(t_i t_j)} \mathrm{Cov}(t_i,t_j)\,[\Delta y_{ij}^2] \quad (7)$$

$$= -\sum_{i<j}^{n}\sum^{n} \frac{\mathrm{Cov}(t_i,t_j)}{E(t_i t_j)}\,[\Delta y_{ij}^2] \quad . \quad (8)$$

Only in special cases, such as with simple random sampling, are expressions (6) and (7) equivalent. It should be emphasized that (6) and (7) (or (8)) are unbiased estimators for $V(Y)$ whenever $E(t_i t_j)>0$ for all distinct pairs of population units. If $E(t_i t_j)=0$ for any pair of units, then special assumptions are needed for an unbiased variance estimator to exist.

E. Remark Concerning the Fixed Sample Size n. It should be clear from the above proofs that (3), (4), and (6) are valid for both fixed and random sample sizes, while (5) and (8) do require a fixed sample size, as assumed. Thus, some of the above theory is more general than initially stated.

F. The Stability of the Variance Estimators. When sampling is without replacement, $v(\hat{Y})$ becomes the familiar Horvitz-Thompson [6] variance estimator and $v_2(\hat{Y})$ becomes the well-known Yates-Grundy [9] estimator. In this case, $v_2(\hat{Y})$ is generally the preferred estimator for $V(\hat{Y})$ because it usually is much more stable than $v_1(\hat{Y})$ and assumes negative values less often. Thus, it would seem reasonable to prefer $v_2(\hat{Y})$ over $v_1(\hat{Y})$ in the general scheme. The sampling variance of $v_2(\hat{Y})$ is quite cumbersome; however, when n=2, as is often the case in practice, $v_2(\hat{Y})$ involves only two sample units and the variance of $v_2(\hat{Y})$ is conveniently obtained from (8) as

$$V[v_2(\hat{Y})] = \sum_{i<j}^{N}\sum^{N} \frac{[\mathrm{Cov}(t_i,t_j)]^2}{E(t_i t_j)} (\Delta y_{ij}^4) - [V(\hat{Y})]^2. \quad (9)$$

G. Remark Concerning Multi-Stage Sampling. This section concludes with one final remark. It should be pointed out that the general theory just developed is applicable in two quite different situations. First of all, the general results are obviously valid when dealing with a single stage sample design. In addition, the general theory is also applicable, without modification, for any multi-stage sampling scheme, as long as the sample size at the final stage is fixed. For example, in a multi-stage design, the y_i are the variate values of the final stage units, and n is the fixed number of these final stage units selected for sample. Of course, for computational reasons and because we often wish to know the variances at the various stages, alternative forms for the variance and its estimators showing the several stages of sampling would have to be developed as needed.

II. A USEFUL APPLICATION OF THE GENERAL THEORY

While it's true that nearly all sample designs in practice are either with or without replacement designs, there does exist at least one ongoing sample survey for which the general theory is quite helpful. The Current Population Survey (CPS) as designed by the U. S. Census Bureau provides us with a useful application of the general theory. The CPS [8] is a stratified multi-stage general population survey of the nation. For simplicity we will focus only on the first stage of selection (for the non-certainty primary units, of course) as if it were the only stage of sampling.

A. The CPS First Stage Sampling Scheme. The sampling scheme used at the first stage of the CPS is certainly an unusual one, and actually resulted from combining two separate existing surveys. We can best describe the sampling scheme for the combined survey as follows. In a typical pair of strata (there are many stratum pairs), choose a sample of size n=3 by initially selecting one stratum at random (i.e., $p=\frac{1}{2}$) and choosing two units with replacement from the chosen stratum using probabilities proportional to some measure of size. Then select one unit with probability proportional to size from the remaining stratum.

B. A Model for Computing the Desired Marginal Probabilities and Expectations. We will now apply the general theory in a typical pair of strata. Let S_1 and S_2 denote the collection of units in stratum 1 and 2, respectively. For simplicity, assume the first N_1 units are in S_1 and the last N_2 units are in S_2 ($N=N_1+N_2$). Let X_h and Y_h (h=1,2) be the stratum totals of the known and unknown variates, respectively. Then,

$$X_h = \sum_{i\in S_h}^{N} x_i \text{ and } Y_h = \sum_{i\in S_h}^{N} y_i \quad (h=1,2).$$

If the i^{th} unit is in stratum h, define the within stratum selection probabilities as

$$p_i' = \frac{x_i}{X_h} \quad (i\in S_h) \text{ for } h=1,2.$$

One simple way to view the selection scheme is to imagine three independent draws from the N units, with one unit selected at each draw. The following three vectors of selection probabilities are used at the various draws:

$$\text{Draw 1: } (p_1', p_2', \ldots, p_{N_1}', 0, 0, \ldots, 0)$$

$$\text{Draw 2: } (0, 0, \ldots, 0, p_{N_1+1}', p_{N_1+2}', \ldots, p_N')$$

$$\text{Draw 3: } \left(\frac{p_1'}{2}, \frac{p_2'}{2}, \ldots, \frac{p_{N_1}'}{2}, \frac{p_{N_1+1}'}{2}, \frac{p_{N_1+2}'}{2}, \ldots, \frac{p_N'}{2}\right)$$

This or any other equivalent model of our sampling scheme allows us to easily compute the following marginal probabilities. For all i we have,

$$\Pr(t_i=1)=\frac{3}{2}p_i' - (p_i')^2, \text{ and } \Pr(t_i=2)=\tfrac{1}{2}(p_i')^2.$$

If the units i and j, $i\neq j$, are in the same stratum, we have

$$\Pr(t_i=1,\ t_j=1) = p_i' p_j' ,$$

while if units i and j are in different strata

$$\Pr(t_i=1,t_j=1) = p_i' p_j'(1-p_i') + p_i' p_j'(1-p_j'), \text{ and}$$

$$\Pr(t_i=1,t_j=2) = \tfrac{1}{2} p_i'(p_j')^2.$$

Using these probabilities the needed expectations are then easily arrived at.

$$E(t_i) = \frac{3}{2} p_i' \quad \text{(all i)},$$

$$E(t_i t_j) = \begin{cases} p_i'\left(\frac{3}{2} + p_i'\right) & i=j \\ p_i' p_j' & \text{i and j in same stratum } (i\neq j) \\ 2p_i'p_j' & \text{i and j in different stratum,} \end{cases}$$

and

$$\mathrm{Cov}(t_i,t_j) = \begin{cases} \frac{p_i'}{2}\left(3-\frac{5}{2} p_i'\right) & i=j \\ -\frac{5}{4} p_i'p_j' & \text{i and j in same stratum } (i\neq j) \\ -\frac{1}{4} p_i'p_j' & \text{i and j in different strata.} \end{cases}$$

These expectations can now be used in conjunction with the general results to derive explicit formulae for $\hat{Y}$, $V(\hat{Y})$, and $v_2(\hat{Y})$.

III. A NEW COMPROMISE SELECTION METHOD FOR n=2 SAMPLE UNITS PER STRATUM

We now turn to a somewhat unrelated topic concerning efficient survey design. One of the simplest techniques for reducing the variance of an estimator is through effective stratification or universe partitioning. Frequently, due to the large amount of auxiliary information available, stratification may be so effective that it is only necessary to select one sample unit per stratum. However, as is well known, samples of size one generally permit only a positively biased estimate of the variance. Consequently, if there is a pressing need for an unbiased variance estimator, the sampler generally redefines his strata by pairing existing strata and selecting a sample of size two from each new stratum pair. If the sample within each new stratum is chosen in such a way that <u>all</u> pairs of distinct universe units have a positive joint probability of occurrence into the sample, then an unbiased estimate of variance will exist. Unfortunately, there is generally a loss in the actual precision obtained by the latter selection method when compared to the former. Appropriately, this decrease in precision associated with the latter method can sometimes be expressed as a simple function of the bias in the variance estimator used with the former method.

This paper shortly proposes a new selection method for the within stratum sample size n=2. This selection scheme is motivated by the frequent need for an unbiased and stable estimate of the variance of $\hat{Y}$, while at the same time sacrificing as little as possible in the actual sampling variance of $\hat{Y}$, thus resulting in an accurate interval estimate for Y. The proposed method is a simple compromise between a stratified scheme where one unit is selected from each of two strata and, the well-known Durbin [4] selection scheme where two units are selected ignoring stratum boundaries. Two unbiased estimators for Y will be proposed and evaluated, along with their unbiased variance estimators.

A. <u>Stratified Scheme - Scheme 1</u>. The stratified selection scheme will be referred to as Scheme 1. In Scheme 1, the within stratum probabilities, $p_i' = x_i / X_h$ ($i\epsilon S_h$, h=1 or 2), are used in the selection of the two sample units, one from each of the two strata. Denote by $\hat{Y}_s$ the usual unbiased estimator for Y using Scheme 1. Using the earlier results, $\hat{Y}_s$ is given by

$$\hat{Y}_s = \sum_{i}^{N} y_i \frac{t_i}{p_i'} = \sum_{h}^{2} \sum_{i\epsilon S_h}^{N} y_i \frac{t_i}{p_i'} \tag{10}$$

with variance

$$V(\hat{Y}_s) = \sum_{h}^{2} \sum_{\substack{i<j \\ S_h}}^{N} \sum^{N} p_i'p_j'\left(\frac{y_i}{p_i'} - \frac{y_j}{p_j'}\right)^2 . \tag{11}$$

Although Scheme 1 provides us with a precise point estimate of Y, the biased interval estimate it also provides may be unacceptable in certain applications.

1. <u>Special Techniques for Estimating the Variance in Scheme 1</u>. Since $E(t_i t_j)=0$ for all distinct pairs in the same stratum, no unbiased variance estimator exists. A biased, usually positively biased, estimate of variance is obtainable by pairing or collapsing the two strata. Several interesting relationships between the bias in the estimate of variance, the actual variance, and the variance that would have been obtained if a sample of size n=2 had been selected from the N units with replacement will now be developed. Let $\hat{Y}_W$ be the estimator for Y using this with replacement scheme. Applying the general theory yields

$$\hat{Y}_W = \sum_{i}^{N} y_i \frac{t_i}{2p_i} \tag{12}$$

with variance

$$V(\hat{Y}_W) = \sum_{i<j}^{N} \sum^{N} 2p_ip_j\left(\frac{y_i}{2p_i} - \frac{y_j}{2p_j}\right)^2, \tag{13}$$

which upon using (11) becomes

$$V(\hat{Y}_W) = \tfrac{1}{2} V(\hat{Y}_s) + \sum_{\substack{i \\ S_1}}^{N} \sum_{\substack{j \\ S_2}}^{N} 2p_ip_j\left(\frac{y_i}{2p_i} - \frac{y_j}{2p_j}\right)^2 . \tag{14}$$

Suppose the i^{th} unit is selected from stratum 1 and the j^{th} unit is selected from stratum 2. Then $Y_s = \frac{y_i}{p_i'} + \frac{y_j}{p_j'}$. An estimator for $V(\hat{Y}_s)$ that is often used is

$$v(\hat{Y}_s;a_1,a_2) = \left(a_1\frac{y_i}{p_i'} - a_2\frac{y_j}{p_j'}\right)^2 , \tag{15}$$

where a_1 and a_2 are known constants and are not dependent upon the two units selected for sample. The expectation of $v(\hat{Y}_s;a_1,a_2)$ is

$$Ev(\hat{Y}_s;a_1,a_2) = \sum_{\substack{i \\ S_1}}^{N} \sum_{\substack{j \\ S_2}}^{N} 4p_ip_j\left(a_1\sqrt{\frac{X_1}{X_2}}\frac{y_i}{2p_i} - a_2\sqrt{\frac{X_2}{X_1}}\frac{y_j}{2p_j}\right)^2 \tag{16}$$

and the expectation of its square is

$$E[v(\hat{Y}_s;a_1,a_2)]^2 = \frac{4X_1X_2}{X^2} \sum_{S_i}^{N} \sum_{S_2}^{N} 4p_ip_j$$

$$\times \left(a_1 \sqrt{\frac{X_1}{X_2}} \frac{y_i}{2p_i} - a_2 \sqrt{\frac{X_2}{X_1}} \frac{y_j}{2p_j} \right)^4 . \quad (17)$$

Let us agree to choose a_1 and a_2 such that $a_1\sqrt{\frac{X_1}{X_2}} = a_2\sqrt{\frac{X_2}{X_1}} = K$. Then using (14), (16) becomes

$$Ev(\hat{Y}_s;a_1,a_2) = 2K^2[V(\hat{Y}_W) - \tfrac{1}{2}V(\hat{Y}_s)], \quad (18)$$

thus showing the bias alluded to earlier, and (17) becomes

$$E[v(\hat{Y}_s;a_1,a_2)]^2 = K^4 \frac{4X_1X_2}{X^2} \sum_{S_1}^{N} \sum_{S_2}^{N} 4p_ip_j$$

$$\times \left(\frac{y_i}{2p_i} - \frac{y_j}{2p_j} \right)^4 . \quad (19)$$

It would be desirable to choose K so that the mean squared error of $v(\hat{Y}_s;a_1,a_2)$, $M[v(\hat{Y}_s;a_1,a_2)] = V[v(\hat{Y}_s;a_1,a_2)] + [Ev(\hat{Y}_s;a_1,a_2)-V(\hat{Y}_s)]^2$, is small. There are three sets of values for a_1,a_2, and K sometimes used in practice.

(i) $a_1 = \sqrt{\frac{X_2}{X_1}}$, $a_2 = \sqrt{\frac{X_1}{X_2}}$, and $K^2=1$, in which case

$$Ev(\hat{Y}_s;a_1,a_2) - V(\hat{Y}_s) = 2[V(\hat{Y}_W)-V(\hat{Y}_s)], \quad (20)$$

that is, the bias is equal to twice the (probable) reduction in the actual variance between the two schemes.

(ii) $a_1 = \frac{X}{2X_1}$, $a_2 = \frac{X}{2X_2}$, and $K^2 = \frac{X^2}{4X_1X_2}$. Since $K^2 \geq 1$, this choice of a_1 and a_2 generally gives a larger bias than does choice 1.

(iii) $a_1 = \frac{2X_2}{X}$, $a_2 = \frac{2X_1}{X}$, and $K^2 = \frac{4X_1X_2}{X^2}$. Since $K^2 \leq 1$ this choice of a_1 and a_2 generally gives a smaller bias than does choice 1.

In the past, the Bureau has frequently used both the first and third sets of "a" weights (a_1,a_2) as given above.

B. Durbin Scheme - Scheme 2. The Durbin [4] selection scheme will be referred to as Scheme 2. In Scheme 2, the basic selection probabilities, $p_i=x_i/X$, are used in conjunction with the Durbin selection method in selecting two sample units from the two combined strata, completely ignoring the stratum boundaries. The Durbin selection scheme is a simple unequal probability without replacement selection scheme that selects n=2 units per stratum, with inclusion probabilities $\pi_i=2p_i$ and joint inclusion probabilities.

$$\pi_{ij} = \frac{2p_ip_j}{\lambda}\left[\frac{1}{1-2p_j} + \frac{1}{1-2p_i}\right], \quad (i\neq j) \quad (21)$$

where $\lambda = 1 + \sum_{k}^{N} \frac{p_k}{1-2p_k}$. The Durbin method of selection has been shown [1,4] to possess several highly desirable properties. This scheme is used at various stages of selection in several surveys at the Bureau.

Let $\hat{Y}_\pi$ be the usual unbiased estimator for Y obtained from Scheme 2. The previous results show that $\hat{Y}_\pi$ is given by

$$\hat{Y}_\pi = \sum^{2} \frac{y_i}{\pi_i} \quad (22)$$

with variance

$$V(\hat{Y}_\pi) = \sum_{i<j}^{N} \sum^{N} (\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 , \quad (23)$$

with π_i and π_{ij} as given above.

C. The New Compromise Scheme—Scheme 3. The new compromise selection method, referred to as Scheme 3, will now be given. This new scheme is a simple combination of Schemes 1 and 2, and is motivated by the desire for a selection scheme that possesses most of the optimum properties of these two schemes. Specifically, we desire the (expected) lower variance associated with Scheme 1 and the unbiased and stable variance estimator accompanying Scheme 2.

Let p be any constant satisfying $0\leq p\leq 1$. Then to apply the new compromise Scheme 3, simply choose either Scheme 1 or Scheme 2 with probabilities p and 1-p, respectively, and proceed to select the sample according to the chosen scheme. In an actual survey situation, of course, Scheme 3 would be applied separately in each of many stratum-pairs. Using Scheme 3, two unbiased estimators for Y will be considered along with the variance and unbiased estimate of variance for each. The properties of these estimators and the considerations involved in the choice of p will be the subject of the remainder of this section.

D. Overall Inclusion Probabilities for Scheme 3. Recall that $\pi_i=2p_i=2x_i/X$ and π_{ij} are the Scheme 2 (Durbin) inclusion and joint inclusion probabilities, respectively, and that $p_i' = \frac{x_i}{X_h}$ $(i\in S_h)$ are the Scheme 1 inclusion probabilities. Then, if unit i is in stratum h (either h=1 or h=2 for every i), the Scheme 3 inclusion probability is the function of p given by

$$\pi_i(p) = (p_i')p + (\pi_i)(1-p) \quad (i\in S_h, h=1 \text{ or } 2). \quad (24)$$

The Scheme 3 joint inclusion probability for units i and j $(i\neq j)$ is the following function of p:

$$\pi_{ij}(p) = \begin{cases} (p_i'p_j')p + (\pi_{ij})(1-p) & \text{if i,j in different strata} \\ (\pi_{ij})(1-p) & \text{if i,j in same stratum.} \end{cases} \quad (25)$$

Since the Durbin method satisfies $\pi_i\pi_j>\pi_{ij}$ for all $i\neq j$, then if units i and j are in different strata, $\pi_{ij}(p)>\pi_{ij}$. Therefore, it is clear that the effect of Scheme 3, when compared to Scheme 2, is to increase the joint occurrence of units in different strata, while decreasing the joint probability of units in the same stratum.

E. Unconditional Estimator for $Y-\hat{Y}_p$. Under Scheme 3, an unconditional estimator for Y is the usual unbiased estimator, which is p dependent,

and is given by

$$\hat{Y}_p = \sum^{N} (t_i) \frac{y_i}{\pi_i(p)} \qquad (26)$$

with variance

$$V(\hat{Y}_p) = \sum_{i<j}^{N\,N} d_{ij}(p)\,[\Delta y_{ij}(p)]^2 \qquad (27)$$

where $d_{ij}(p) = \pi_i(p)\pi_j(p) - \pi_{ij}(p)$, and

$$\Delta y_{ij}(p) = \left(\frac{y_i}{\pi_i(p)} - \frac{y_j}{\pi_j(p)}\right).$$

Although probably not obvious from (27), $V(\hat{Y}_p)$ is not necessarily monotone (decreasing or increasing) between p=0 and p=1. As we will soon see in the numerical examples, $V(\hat{Y}_p)$ can either be monotone or have peaks and valleys between the two endpoints p=0 and p=1. Thus, one should have sufficient information in order to efficiently specify a value of p when applying Scheme 3.

The unbiased Yates-Grundy estimator for $V(\hat{Y}_p)$ is

$$v(\hat{Y}_p) = \frac{d_{ij}(p)}{\pi_{ij}(p)}\,[\Delta y_{ij}(p)]^2 \,, \quad (p\neq 1) \qquad (28)$$

where the i^{th} and j^{th} units are the selected units.

As can be seen from both the variance estimator $v(\hat{Y}_p)$ in (28) and its variance, $V[v(\hat{Y}_p)]$ obtained obtained from (9), the stability of our variance estimator is dependent upon both p and the effectiveness of the stratification. Although we can't allow p to become too large (near unity), one would expect this scheme can tolerate larger values of p, if desirable, when stratification is effective than if it is not. This is because, although $\pi_{ij}(p)$ is small for units in the same stratum, so also is $[\Delta y_{ij}(p)]^2$ whenever stratification is effective.

In summary, to efficiently apply the unconditional estimator $\hat{Y}_p$ under Scheme 3, one must attempt to find a value of p that jointly produces a small true variance for the estimator of Y, and a stable variance estimator. The criterion used in this paper to quantify the preceding sentence is to find the value of p that minimizes

$$Q_p = V(\hat{Y}_p) + \sqrt{V[v(\hat{Y})_p]} \quad (p\neq 1). \qquad (29)$$

A small value of Q_p should, in some sense, tend to indicate a "good" interval estimate for Y, on the average. Other possible measures of the accuracy of our interval estimate would include differentially weighting each of Q_p's components. The requirements of the survey and the statistician's subjective and objective judgments would ultimately determine these weights.

F. Conditional Estimator for $Y-\hat{Y}_c$. Under Scheme 3, a conditionally (conditioned on the randomly selected scheme) unbiased estimator for Y is given by

$$\hat{Y}_c = \begin{cases} \hat{Y}_s & \text{if Scheme 1 is chosen} \\ \hat{Y}_\pi & \text{if Scheme 2 is chosen} \end{cases} \qquad (30)$$

with variance

$$V_p(\hat{Y}_c) = p\,V(\hat{Y}_s) + (1-p)\,V(\hat{Y}_\pi) \qquad (31)$$

$$= \sum_{i<j}^{N\,N} [\pi_i\pi_j - \alpha_{ij}(p)]\Delta y_{ij}^2 \qquad (32)$$

where $\alpha_{ij}(p)$ is defined by

$$\alpha_{ij}(p) = \begin{cases} \pi_i\pi_j p + (\pi_{ij})(1-p) & \text{if units i,j are in different strata} \\ (\pi_{ij})(1-p) & \text{if units } i\neq j \text{ are in same stratum} \end{cases} \qquad (33)$$

and where $\Delta y_{ij} = \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)$. Note that although $\hat{Y}_c$ does not depend upon p, its sampling distribution does. It is obvious from (31) that, unlike $V(\hat{Y}_p)$, $V_p(\hat{Y}_c)$ is monotone between p=0 and p=1, and further, $V(\hat{Y}_\pi)$ and $V(\hat{Y}_s)$ uniquely determine $V_p(\hat{Y}_c)$. Thus, if stratification is effective, it follows that for all p,

$V_1(\hat{Y}_c) = V(\hat{Y}_s) \leq V_p(\hat{Y}_c) \leq V(\hat{Y}_\pi) = V_0(\hat{Y}_c)$. The unbiased Yates-Grundy type estimator for $V_p(\hat{Y}_c)$ is

$$v_p(\hat{Y}_c) = \frac{\pi_i\pi_j - \alpha_{ij}(p)}{\pi_{ij}(p)}\,[\Delta y_{ij}^2] \quad (p\neq 1) \qquad (34)$$

The comments just made concerning the stability of $v(Y_p)$ hold here for $v_p(Y_c)$ also.

Therefore, when stratification is effective, one should choose p as large as possible, subject to the constraint of a stable variance estimator. The suggested criterion, analogous to the earlier one, is to choose p such that

$$Q_{cp} = V_p(\hat{Y}_c) + \sqrt{V[v_p(\hat{Y}_c)]} \quad (p\neq 1) \qquad (35)$$

is minimized. This optimum value of p should then provide us with an accurate interval estimate for Y.

Scheme 1 is clearly a special case of Scheme 3 and is obtained by simply letting p=1. For this case, both the conditional and the unconditional Scheme 3 estimators become equivalent to the stratified estimator $\hat{Y}_s$, and thus, our criterion for measuing the accuracy of the interval estimates becomes

$$Q_s = V(\hat{Y}_s) + \sqrt{M[v(\hat{Y}_s;a_1,a_2)]} \,, \qquad (36)$$

and is dependent upon the "a" weights chosen.

IV. TWO NUMERICAL EXAMPLES USING HORVITZ & THOMPSON'S NATURAL POPULATION

In this final section, two numerical examples are considered. For each illustration, the properties of Schemes 1,2, and 3 are explored. As the examples show, the performance of any of the schemes significantly depend upon the population and the quality of the stratification. The first example demonstrates the significant gains obtained by effective stratification, the associated overestimation of the variance, and how Scheme 3 can serve as an effective compromise between Schemes 1 and 2. The second example is included to demonstrate the consequences of ineffective stratification.

A. Horvitz and Thompson's Natural Population. In their 1952 paper, Horvitz and Thompson [6] investigated a universe consisting of N=20 blocks in Ames, Iowa. The data is given in table 1, where the measures x_i are the number of eye-estimated households on the i^{th} block and the y_i are the actual number of households. The data has

been reordered here for clarity. Many authors have subsequently tested their sampling schemes on this population. Table 2 is a summary of results obtained by Horvitz and Thompson [6], Hartley and Rao [5], and Raj [7]. Two numerical examples dealing with Scheme 3 will be given.

Table 1

HORVITZ-THOMPSON NATURAL POPULATION

i	y_i	x_i	y_i/x_i
1	19	18	1.06
2	9	9	1.00
3	21	24	.88
4	22	25	.88
5	15	14	1.07
6	18	18	1.00
7	37	40	.93
8	12	12	1.00
9	27	27	1.00
10	25	26	.96
11	19	19	1.00
12	12	12	1.00
13	17	14	1.21
14	14	12	1.17
15	27	23	1.17
16	20	17	1.18
17	25	21	1.19
18	35	24	1.46
19	47	30	1.57
20	13	9	1.44
	Y=434	X=394	

Table 2

SUMMARY OF PREVIOUS RESULTS

Sampling Scheme	Variance of the Estimator	$\left(\text{Variance of Variance Estimator}\right)^{\frac{1}{2}}$
1. Simple Random Sampling	17,122	26,539
2. Stratified Random; one element from each of two strata with equal probability[1]	7,873	NA
3. Equal Probability Systematic Sampling	10,224	NA
4. pps With Replacement	3,247	4,611
5. First Horvitz-Thompson Scheme (πps)[2]	3,095	NA
6. Second Horvitz-Thompson Scheme (πps)[3]	3,075	NA
7. Systematic πps	3,014	3,983

NA = Not Available

[1]Stratum 1 consists of the 10 blocks with the largest measures of size ($x_i \geq 19$), with the smallest ($x_i \leq 18$) 10 blocks in stratum 2.

[2]First sample unit selected with pps, second unit from remainder with equal probabilities. Original measures altered so as to obtain an approximate πps (i.e., $\pi_i = 2x_{i/X}$) scheme.

[3]First sample unit selected with pps, second unit with pps of the remaining units. Original measures altered so as to obtain an approximate πps scheme.

B. Example 1-First Stratification. In the first example, units 1 through 12 comprise stratum 1 and units 13 through 20 comprise stratum 2. From table 1 this would appear to be an effective stratum formation. We have $N_1=12$, $N_2=8$, $X_1=244$, $X_2=150$, $Y_1=236$, and $Y_2=198$.

The results appear in tables 3a and 3b and are encouraging. In this example, the unconditional estimator yields excellent point and interval estimates for any p satisfying $.25 \leq p \leq .65$. In this case the precision obtained compares well with that of Scheme 1. We also see that $V(Y_p)$ behaves quite smoothly in this first stratification. As shown in table 3b, the bias in each of the Scheme 1 variance estimators is probably intolerable to most. In fact, this bias is so sizeable that for nearly each value of p not too near unity, both the unconditional and the conditional estimators are superior to the Scheme 1 estimator when applying the "Q" criterion. Finally, for each p, the unconditional estimator is always superior to the conditional estimator.

C. Example 2-Second Stratification. The effectiveness of the stratification is an important issue, as this second and final example demonstrates. Horvitz and Thompson suggest stratifying according to the measures of size (x_i), with the 10 largest eye-estimated blocks in stratum 1 and the 10 smallest in stratum 2. When sampling is with equal probabilities, this method of stratification has already been considered, as shown in table 2 (No. 2). In this case, there was a significant improvement compared to unrestricted simple random sampling (table 2, No. 1). As we will see, such is not the case when comparing stratified unequal probability sampling (using the strata definition just given) with (unrestricted) pps with replacement sampling (table 2, No. 4). Thus, in this example, all units with $x_i \geq 19$ (i=3,4,7,9,10,11,15,17,18 and 19) are defined as stratum 1, and all units with $x_i \leq 18$ (i=1,2,5,6,8,12,13,14,16, and 20) comprise stratum 2. The summary totals are now $N_1=N_2=10$, $X_1=259$, $X_2=135$, $Y_1=285$, and $Y_2=149$.

Inspection of the y_{i/x_i} column in table 1 tend to indicate this second stratification is not very effective. The stratified scheme yields considerably less precision ($V(Y_s)=4025$) than either the Durbin scheme or pps with replacement sampling. Because the stratification was so poor, the precision of both the conditional and the unconditional estimators get steadily worse as $0 \to p \to 1$, although the unconditional estimator begins to dip back down at about $p \doteq .75$. The precision of both variance estimators become steadily worse as $0 \to p \to 1$ because the small joint inclusion probabilities are not being associated with small $[\Delta y_{ij}(P)]^2$, again due to poor stratifying. In addition, each of the three Scheme 1 variance estimators seriously underestimates (2438, 2739, and 2224) the actual variance, whereas when stratification is effective they are generally each overestimates of variance. Therefore, as this example indicates, the quality (or lack of quality) of the stratification is a crucial issue, and, in particular, stratifying only on the basis of size is certainly questionable. The tables showing the analysis for this second example can be obtained upon writing the author.

TABLE 3a

Horvitz-Thompson Population - First Stratification Results						
	Unconditional Estimator			Conditional Estimator		
Scheme 3	$V(\hat{Y}_p)$	$\sqrt{V[v(\hat{Y}_p)]}$	Q_p	$V_p(\hat{Y}_c)$	$\sqrt{V[v_p(\hat{Y}_c)]}$	Q_{cp}
p=0 (Durbin, Scheme 2)	3011	3990	7001	3011	3990	7001
p=.10	2220	2809	5029	2791	3382	6173
p=.25	1438	2042	3480	2463	2717	5180
p=.50	896	2544	3440	1915	2516	4431
p=.65	842	3356	4198	1586	3129	4715
p=.75	853	4202	5055	1367	3963	5330
p=.90	863	7147	8010	1038	6992	8030
p=1 (Stratified, Scheme 1)	819	--	--	819	--	--

TABLE 3b

Horvitz-Thompson Population - First Stratification Results					
	a_1	a_2	$Ev(\hat{Y}_s;a_1,a_2)$	$\sqrt{M[v(\hat{Y}_s;a_1,a_2)]}$	Q_s
1.	$\sqrt{X_{2/X_1}}$	$\sqrt{X_{1/X_2}}$	5674	6984	7803
2.	$X/_{2X_1}$	$X/_{2X_2}$	6017	7440	8259
3.	$2X_{2/X}$	$2X_{1/X}$	5351	6554	7373

REFERENCES

1. C. Asok and B. V. Sukhatme. On Sampford's Procedure of Unequal Probability Sampling Without Replacement. Journal of the American Statistical Association, (1976), Vol. 71, pp. 912-918.
2. W. G. Cochran. Sampling Techniques. 2nd ed. New York: Wiley and Sons, 1963.
3. J. Cornfield. On Samples from Finite Populations. Journal of the American Statistical Association, (1944), Vol. 39, pp. 236-239.
4. J. Durbin. Design of Multi-stage Surveys for the Estimation of Sampling Errors. Applied Statistics, (1967), Vol. 16, pp. 152-164.
5. H. O. Hartley and J. N. K. Rao. Sampling With Unequal Probabilities and Without Replacement. Annals of Mathematical Statistics, (1962), Vol. 33, pp. 350-374.
6. D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association, (1952) Vol. 47, pp. 663-685.
7. D. Raj. Sampling Theory. 1st ed. New York: McGraw-Hill, 1968.
8. M. Thompson and G. Shapiro. The Current Population Survey: An Overview. Annals of Economic and Social Measurement, (1973), Vol. 2, No. 2.
9. F. Yates and P. M. Grundy. Selection Without Replacement From Within Strata With Probability Proportional to Size. Journal of the Royal Statistical Society, (1953), Series B, Vol. 15, pp. 253-261.

ACKNOWLEDGEMENTS

The author would like to thank Gary Sparks for his excellent computer programming, Kirk Wolter for some helpful comments, and Edith Oechsler for her conscientious typing, all of the Census Bureau. Thanks also to Dr. Harry Rosenblatt of the American University, Washington, D. C., for reviewing a larger version of this paper.

NUMERICAL COMPARISON OF SOME ESTIMATORS OF VARIANCE UNDER PPS SYSTEMATIC SAMPLING

Cary T. Isaki and Susan J. Pinciaro, Bureau of the Census

I. Introduction

The purpose of this study is to investigate the performance of several estimators of the variance of the Horvitz-Thompson (HT) estimator of total, $\hat{Y}_{HT} = \sum_{i=1}^{n} y_i/\pi_i$, under a probability proportional to size (PPS) systematic sampling design. The PPS systematic sampling scheme was selected for study because of its wide applicability and usage. Since the PPS systematic scheme does not yield an unbiased estimator of variance, a comparative study of the biases and mean square errors (MSE's) of several variance estimators in a real finite population was conducted.

The population used in the study consisted of mobile home dealers canvassed in the 1972 Census of Retail Trade. The estimators of variance chosen for study include those most commonly found in the literature plus some minor variations. We considered two variables, referred to as y and z, as characteristics to be estimated. All of the results were obtained for two universes distinguished by two different orderings of the population of mobile home dealers. We refer to the population ordered by decreasing measure of size as Universe I and to a second ordering of the population (roughly a geographical ordering of the units) as Universe II. Given the results from the two universes, the effect that the order of the units in the frame has upon the variance of the estimator of total and on the estimation of variance is considered.

II. Description of the Study

1. Preparation

The population used in the study consisted of a compact file of mobile home dealers canvassed in the 1972 Census of Retail Trade. The data record for each mobile home dealer contained an identification number, 1972 annual sales, 1972 average quarterly payroll and 1972 first-quarter employment. Universe II was obtained by sorting on the identification number. For the purposes of this study, the 19 largest mobile home dealers were excluded on the basis of their size (these units would be designated as certainty units in most sample designs), and a few of the very smallest dealers (in terms of payroll) were excluded to simplify the computer programming. We considered 1972 annual sales (y) and 1972 first-quarter employment (z) as characteristics to be estimated and 1972 average quarterly payroll (x) was used as a measure of size, i.e., $p_i = x_i/X$ and $X = \sum_{i=1}^{N} x_i$. The payroll figures for a few of the dealers were adjusted slightly so that X was divisible by the sample sizes (n=30, 60, 150, 300) considered in the study.

2. Estimators of Variance

The following variance estimators were utilized for estimating the variance of $\hat{Y}_{HT}$ and the variance of $\hat{Z}_{HT} = \sum_{i=1}^{n} z_i/\pi_i$ over all possible systematic samples of sizes n=30, 60, 150, and 300 from each of the two universes:

a. Random group estimator with t groups -

$$rg(t) = 1/t \sum_{g=1}^{t} \frac{(y_g - \bar{y})^2}{t-1} \qquad t=5,10,15,20,30$$

b. With replacement variance estimator -

$$WR = \sum_{i=1}^{n} \frac{\left(\frac{y_i}{p_i} - \hat{Y}_{HT}\right)^2}{n(n-1)}$$

c. With replacement variance estimator with adjustment -

$$WRA = \left[1 - \sum_{i=1}^{n} \frac{\pi_i}{n}\right] \times WR$$

d. "Randomized systematic" variance estimator-

$$RRS = \frac{1}{n-1} \sum_{i<i'}^{n} \left\{ 1-(\pi_i + \pi_{i'}) + \sum_{k=1}^{N} \frac{\pi_k^2}{n} \right\} \left(\frac{y_i}{\pi_i} - \frac{y_{i'}}{\pi_{i'}}\right)^2$$

e. Collapsed stratum variance estimator -

$$CS = \sum_{h=1}^{n/2} \left(\frac{y_i}{np_i} - \frac{y_j}{np_j}\right)_h^2$$

where i and j are adjacent pairs of units in the sample.

f. Successive pairs variance estimator -

$$SP = \frac{n}{2(n-1)} \sum_{i=1}^{n-1} \left(\frac{y_i}{np_i} - \frac{y_{i+1}}{np_{i+1}}\right)^2$$

g. Successive pairs variance estimator with adjustment -

$$SPW = \left[1 - \sum_{i=1}^{n} \frac{\pi_i}{n}\right] \times SP$$

All of the variance estimators specified above can either be found in the literature or in the usual sampling texts. The RRS estimator was proposed [4,7] in the context of PPS systematic sampling when the units in the population to be sampled from are randomly arranged in one of the N! possible sequences. Although the population under

study was placed separately in two specified orders, it was felt that it would be of interest to include RRS in the comparison. The CS and SP estimators were felt to be reasonable estimators of the variance under a PPS systematic design when one visualized the actual sample design as being approximated by a one sample unit per stratum design where the strata consist of units lying within the realized sampling intervals. The WR estimator is a special case of rg(t) when n=t. It can be shown that WR has the same bias of rg(t) but a smaller MSE than rg(t).

In addition to the estimators listed above, another estimator which we call the pseudo random group (prg) estimator was considered. Estimators prg(t) and rg(t) have the same form, but they differ in the manner in which the sample units are assigned to the t groups. In rg(t) the sample units are assigned randomly to the t groups while in prg(t) the sample units are assigned to groups systematically in the order which they are selected into the sample.

$\hat{Y}_{HT}$ and $\hat{Z}_{HT}$ were calculated for every possible sample of a given size, the samples of units not necessarily being unique, and $V(\hat{Y}_{HT})$ and $V(\hat{Z}_{HT})$ were calculated for each sample size in each universe. The expected value of each of the estimators of $V(\hat{Y}_{HT})$ and $V(\hat{Z}_{HT})$ (except rg(t)) was obtained by averaging the estimates over all possible samples of the given size. The variance of each of these variance estimators was also calculated.

The mean and variance of rg(t) were calculated in the following manner.

i. Using the result referred to earlier, we set

$$E[rg(t)] = E[WR]$$

ii. It can be shown that

$$Var[rg] = E\{Var[rg|sample]\} + Var[WR].$$

Hence Var[rg|sample] was calculated for each sample and averaged over all samples. This term was then added to Var[WR].

Having obtained the mean and variance of each estimator, we calculated the mean square error of each. The results of these calculations are provided in Tables 1 and 2.

Further distributional properties of the estimators of $V(\hat{Y}_{HT})$ are reflected by the confidence interval results in Table 3. These proportions were obtained in the following manner. For a given sample, 90 and 95% confidence intervals were constructed for Y (and Z) using the $\hat{Y}_{HT}$ ($\hat{Z}_{HT}$) estimate and each of the estimates of ($\hat{V}_{HT}$)($V(\hat{Z}_{HT})$) produced by that sample. For example, for 95% confidence intervals, $\hat{Y}_{HT} \pm 1.96\sqrt{SPY}$ was calculated for each sample where SPY = the SP estimator of $V(\hat{Y}_{HT})$ for the given sample. 90 and 95% confidence intervals using $\hat{Y}_{HT}$($\hat{Z}_{HT}$) and its variance for each possible PPS systematic sample were also constructed (e.g., for 95% confidence intervals, $\hat{Y}_{HT} \pm 1.96\sqrt{V(\hat{Y}_{HT})}$). Then, for each estimator, the true proportion of the confidence intervals which contained Y(Z) was calculated as was the proportion of confidence intervals constructed using the variance of $\hat{Y}_{HT}$ ($\hat{Z}_{HT}$). These calculations were made for each sample size and universe. The proportions are provided in Table 3.

3. Summary Parameters

The intraclass correlation, ρ, was calculated for each sample size and is shown in Table 7 along with its lower bound, $-(\frac{1}{n-1})$. ρ is defined in [3] where it is shown to be expressible alternatively as

$$\rho_Y = \frac{1}{n-1}\,[V(\hat{Y}_{HT}) - V(\hat{Y}')]/V(\hat{Y}')$$

where $V(\hat{Y}')$ is the variance of the estimator $\hat{Y}' = \sum_{i=1}^{n} y_i/np_i$, under with replacement PPS sampling; that is,

$$V(\hat{Y}') = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2 .$$

The term referred to as VS in Table 7 is an approximation to the variance of $\hat{Y}_{HT}$ under a sample design in which the units in the population are randomly ordered and a PPS systematic design is used to select the sample of n units [4,7]. It has the following form:

$$VS = \sum_{i=1}^{N} \pi_i\left[1 - \frac{n-1}{n}\,\pi_i\right]\left(\frac{y_i}{\pi_i} - \frac{Y}{n}\right)^2 .$$

Its magnitude, relative to $V(\hat{Y}_{HT})$, is presented in Table 7. The remaining columns in Table 7 provide ratios of the variances resulting from several alternative estimator-sample design pairs relative to $V(\hat{Y}_{HT})$.

III. Results

1. Estimators of $V(\hat{Y}_{HT})$ and $V(\hat{Z}_{HT})$

For the population of N=5634 mobile home dealers,

$$Y = \sum_{i=1}^{N} y_i = .32385 \times 10^7$$

$$Z = \sum_{i=1}^{N} z_i = .33213 \times 10^5$$

and $$X = \sum_{i=1}^{N} x_i = .57300 \times 10^5$$

Graphs 1 and 2 illustrate the relationships between the variables y and x and between z and x in the population. The plots indicated that x would be a useful design variable. The correlation coefficients squared are, respectively, .74 and .75.

Tables 1A and 1B present, for Universe I, the expected values and MSE's (relative to MSE(WR)) of the estimators discussed in Section II.2. The MSE's of rg(•) for n=150 and 300 were not calculated due to limited resources and because it was observed that the MSE's for rg(•) for n=30 and 60 did not differ appreciably from the MSE's for prg(•). In the case of rg(•), when the sample size was such that the random groups did not contain equal numbers of sample units, the MSE was not calculated. Tables 2.A and 2.B are similar to Tables 1.A and 1.B except that the results refer to Universe II. In the following all conclusions and summaries refer solely to the mobile home dealer population under study.

In terms of relative bias, CS had the smallest bias in the largest number of the 8 characteristic/sample size combinations in Universe I and appeared to possess a bias slightly larger than that of the smallest in the other cases. In Universe II, WR had the smallest relative bias for the y characteristic while for the z characteristic, no estimator stood out.

With respect to MSE, SPW consistently exhibited the smallest MSE in Universe I. Other estimators with reasonably small MSE's were SP, CS, and WRA. For Universe II, RRS, WRA, CS, pg (15) and SPW had the smallest MSE for at least one characteristic/sample size combination with RRS appearing best overall. In general, the estimators rg(•) and prg(•) performed poorest of all over the 16 cases with prg(•) performing better than rg(•).

One interesting observation can be made with respect to WR and RRS and the relative bias. That is, for a given characteristic/sample size combination each of the estimators exhibit, approximately, the same expected value whether applied in Universe I or II. When $\rho_y(\rho_z)$ is negative and hence PPS systematic sampling is superior to PPS with replacement sampling, the relative biases of WR and RRS are positive and vice versa when $\rho_y(\rho_z)$ is positive (except for one case). This result probably occurs because WR and RRS do not reflect the systematic nature of the sampling design as compared to CS and SP. Hence, when ρ is negative, and the ratios y_i/π_i in the sample are diverse, WR estimates too high, and when ρ is positive the ratios in the sample are similar and hence WR estimates too low.

The results of the confidence interval calculations are located in Table 3. The proportions obtained from intervals constructed using $\hat{Y}_{HT}(\hat{Z}_{HT})$ and $V(\hat{Y}_{HT})(V(\hat{Z}_{HT}))$ are, in most of the 16 universe/characteristic/sample size combinations, greater than the .90 (or .95) which would have been expected from a normally distributed $\hat{Y}_{HT}(\hat{Z}_{HT})$. In those cases in which the proportions did not exceed .90 (or .95) they were very close.

A few general comments may be made concerning the proportions resulting from the confidence intervals constructed with $\hat{Y}_{HT}(\hat{Z}_{HT})$ and the estimates of $V(\hat{Z}_{HT})$). As the sample size increased the number of individual proportions which exceeded the .90 (or .95) levels rose. Over both universes, and for sample sizes n=30,60,150,300 the number of proportions greater than .90 (or .95) equaled 10,14,36, and 29, respectively (out of 88). Of these, 9,14,26, and 22 were in Universe I. Very few proportions from Universe II ever reached the .90 (or .95) levels.

In terms of the performances of the individual variance estimators in producing their associated proportions, prg20, prg30, WR, WRA and RRS produced the highest proportions in nearly every universe/characteristic/sample size combination. It was the proportions resulting from these estimators which most often exceeded the .90 (or .95) levels.

3. Alternative Estimator-Sample Design Pairs

Table 7 illustrates the ρ value for each universe/characteristic/sample size combination along with the variances of seven other estimator-sample design pairs. The column headed $V_{SYS}[\cdot]$ represents the variance of $\hat{Y}_{HT}(\hat{Z}_{HT})$ under PPS systematic sampling using quarterly payroll as the measure of size. $V_{WR}[\cdot]$ represents the variance of $\hat{Y}' = \sum_{i=1}^{n} \frac{y_i}{np_i}$ ($\hat{Z}' = \sum_{i=1}^{n} \frac{z_i}{np_i}$) under PPS with replacement sampling. $V_{SRS}[\cdot]$ represents the variance of the HT estimator of total under an SRS without replacement design, and $V_{SRS}[\text{Ratio, X}]$ denotes the variance of the ratio estimator using X = quarterly payroll as the auxiliary variable under an SRS without replacement design. $V_{SYS}[\text{Ratio, Z}]$ refers to the variance of the ratio estimator of total using Z as the auxiliary variable under a PPS systematic design. $V_{WR}[\text{Ratio, Z}]$ represents the variance of the ratio estimator of total using Z as the auxiliary variable under a PPS with replacement design. $V_{SRS}[\text{Ratio, Z}]$ is analogous to $V_{SRS}[\text{Ratio, X}]$ with Z used as the auxiliary variable. The entries in the table express the above-described variances relative to $V_{SYS}(\cdot)$.

The simple raw correlation between Y and Z is $\rho_{Y,Z} = .789$. This implies, since $\rho_{Y,Z} > 1/2 \frac{C.V.(Z)}{C.V.(Y)}$, in the simple random sampling context, that the ratio estimator is preferable. The entries in the table under $V_{SRS}(\cdot)$ and $V_{SRS}[\text{Ratio, Z}]$ support the choice of the ratio estimator in this situation. However, neither of these estimator-sample design pairs does better than $V_{SYS}(\cdot)$ for any universe/characteristic/sample size combination.

Comparison of $V_{SYS}(\cdot)$ and $V_{SYS}[\text{Ratio,Z}]$ shows that the HT estimator performs better in 5 or 8 cases, and the ratio estimator does better in the other three cases. In the PPS systematic context, the relevant correlation in deciding between the HT estimator and a ratio estimator is no longer the raw correlation between Y and Z, but is the correlation between $\hat{Y}_{HT}$ and $\hat{Z}_{HT}$, designated as $\rho_{\hat{Y},\hat{Z}}$, and the criterion for selection of the ratio estimator over the HT estimator

is

$$\rho_{\hat{Y},\hat{Z}} \quad 1/2 \; \frac{C.V.(\hat{Z}_{HT})}{C.V.(\hat{Y}_{HT})}$$

In those cases in which the ratio estimator is superior to HT (has smaller variance), even when accounting for the bias of the ratio estimator, it remains better than HT. This follows from the results of Table 4 where it is seen that the MSE of $\hat{Y}_R$ is lower than the variance of $\hat{Y}_{HT}$ in Universe II for n=60,150 and 300. However, in these cases, the estimator-sample design producing $V_{WR}(\cdot)$ performs even better than V_{SYS}[Ratio, Z] and, from the table we see that VS is even better than $V_{WR}(\cdot)$ in these cases (VS can be shown to be better than $V_{WR}(\cdot)$ in general).

In almost all cases, when $\rho<0$, $V_{SYS}(\cdot) \leq VS \leq V_{WR}(\cdot)$. From [7], we know that $VS \leq \overline{V}_{WR}(\cdot)$; hence, when $\rho>0$, $VS \leq V_{WR}(\cdot) \leq V_{SYS}(\cdot)$. From a practical standpoint, if we decide to use PPS systematic sampling and the Horvitz-Thompson estimator and suspect that $\rho<0$, we can use VS as a "safe" (larger than $V_{SYS}(\cdot)$) approximation to $V_{SYS}(\cdot)$ for design purposes.

In general, Table 7 shows that the estimator-sample design pairs resulting in VS or $V_{SYS}(\cdot)$ are better than the rest, and the sign of ρ appears to determine which of the two is preferable. Also, Table 7 demonstrates that gains of at least 20% in $V_{SYS}(\cdot)$ can be realized by using Universe I over Universe II, a consequence of the negative ρ induced by the ordering.

4. Conclusion

In practice, one never really knows whether ρ will be negative with respect to the characteristics to be estimated. However, in many instances comparable data is available on the same population for a previous point in time. Graphs 3 and 4 are plots of the ratios of sales to payroll and employment size to payroll, respectively. As is evident in the two graphs, an ordering of the units by size of payroll and a systematic sampling scheme will tend to spread the ratios evenly over the possible samples and hence make the ratios within the samples diverse, thereby possibly inducing a negative ρ. It is speculated that the same analysis performed on other populations with similar graphs as those of the mobile home dealer population will produce results comparable to the variance estimator comparisons arrived at as a result of Tables 1, 2 and 3. Hence, faced with another population of interest with similar graphs as in Graphs 3 and 4, one can use the results concerning the variance estimators of this limited study in the decision-making process.

The tables containing the results on the estimation of $Var(\hat{Y}_{HT})$ and $Var(\hat{Z}_{HT})$ in Universe II have been omitted due to space limitations. Also, both text and tables relating to the estimation of $Var(\hat{Y}_R)$, where $\hat{Y}_R=(\hat{Y}_{HT}/\hat{Z}_{HT})\,Z$, have been omitted, as have all graphs referred to in the text. Interested readers may contact the authors for these results.

REFERENCES

[1] Cochran, W.G. (1963), Sampling Techniques, John Wiley and Sons, N.Y.

[2] Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), Sample Survey Methods and Theory, Vol. 1, John Wiley and Sons, N.Y.

[3] Hartley, H.O., "Systematic Sampling With Unequal Probability and Without Replacement," Journal of the American Statistical Association, Vol. 61 (1966), 739-748.

[4] Hartley, H.O. and Rao, J.N.K., "Sampling With Unequal Probabilities and Without Replacement," Annals of Mathematical Statistics, Vol. 33 (1962), 350-374.

[5] Madow, William G. and Madow, Lillian G., "On the Theory of Systematic Sampling, I," Annals of Mathematical Statistics, Vol. 15 (1944), 1-24.

[6] Raj, Des, "Variance Estimation in Randomized Systematic Sampling With Probability Proportional to Size," Journal of the American Statistical Association, Vol. 60 (1965), 278-284.

[7] Rao, J.N.K., "On Three Procedures of Unequal Probability Sampling Without Replacement," Journal of the American Statistical Association, Vol. 58, (1963), 202-215.

[8] Wolter, Kirk M. et al, "Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys," Proceedings of the Business and Economic Statistics Section, American Statistical Association, 1976.

[9] Woodruff, Ralph S., "Simple Method of Approximating Variances of a Complicated Estimate," Journal of the American Statistical Association, Vol. 66 (1971), 411-414.

TABLE 1 A. Expected Value and Relative MSE $\left(\text{Rel MSE} = \frac{\text{MSE}(\cdot)}{\text{MSE(WR)}}\right)$ of Some Variance Estimators of $\hat{Y}_{HT}$ (Universe I)

	n = 30		n = 60		n = 150		n = 300	
	Expected Value x 10^{12}	Rel MSE	Expected Value x 10^{11}	Rel MSE	Expected Value x 10^{11}	Rel MSE	Expected Value x 10^{11}	Rel MSE
prg (5)	.1616	.8502	.7125	.8449	.2879	.8823	.1384	.8067
prg (10)	.1658	.9026	.7970	.8508	.3003	.7168	.1404	.8898
prg (15)	.1741	.9807	.8096	.9624	.3117	1.0084	.1500	.9518
prg (20)	.2525	1.0770	.8228	.8998	.3624	.9412	.1482	.7324
prg (30)	.1815	1.0000	.8650	.9784	.3245	.9597	.1544	.9884
WR	.1815	1.0000	.9039	1.0000	.3609	1.0000	.1803	1.0000
WRA	.1783	.9655	.8718	.9277	.3288	.8164	.1482	.6784
RRS	.1803	.9982	.8923	.9943	.3493	.9743	.1687	.9569
CS	.1468	.8076	.7230	.9701	.2864	.7009	.1424	.6820
SP	.1224	.2691	.6364	.3228	.2769	.7942	.1417	.8060
SPW	.1202	.2603	.6139	.2932	.2522	.6624	.1164	.6146
rg (5)	.1815	1.0433	.9039	1.0951				
rg (10)	.1815	1.0151	.9039	1.0381				
rg (15)	.1815	1.0074	.9039	1.0219				
rg (20)			.9039	1.0145				
rg (30)	.1815	1.0000	.9039	1.0071				
$V(\hat{Y}_{HT})$	.1428		.7177		.2762		.1642	
MSE (WR)	$.2843 \times 10^{24}$		$.3513 \times 10^{23}$		$.2260 \times 10^{22}$		$.2758 \times 10^{21}$	

TABLE 1 B. Expected Value and Relative MSE $\left(\text{Rel MSE} = \frac{\text{MSE}(\cdot)}{\text{MSE(WR)}}\right)$ of Some Variance Estimators of $\hat{Y}_{HT}$ (Universe I)

	n = 30		n = 60		n = 150		n = 300	
	Expected Value x 10^{8}	Rel MSE	Expected Value x 10^{7}	Rel MSE	Expected Value x 10^{7}	Rel MSE	Expected Value x 10^{7}	Rel MSE
prg (5)	.1256	1.1964	.6074	1.4989	.2146	2.1343	.1092	1.5865
prg (10)	.1237	1.0870	.6216	1.2362	.2186	1.4194	.1118	1.4042
prg (15)	.1207	1.0199	.5883	1.0624	.2304	1.3938	.1094	1.0014
prg (20)	.1960	1.5726	.6157	1.0982	.2625	1.5744	.1114	1.0059
prg (30)	.1226	1.0000	.6025	1.0241	.2315	1.0939	.1163	1.1348
WR	.1226	1.0000	.6122	1.0000	.2144	1.0000	.1223	1.0000
WRA	.1204	.9630	.5904	.9179	.2226	.8012	.1005	.4893
RRS	.1214	.9947	.6003	.9839	.2330	.9480	.1110	.7708
CS	.1075	.8996	.5343	.8365	.2086	.8455	.1064	.6500
SP	.0982	.3514	.5052	.3264	.2041	.7280	.1063	.6483
SPW	.0964	.3408	.4872	.3042	.1859	.6442	.0873	.3177
rg (5)	.1226	1.2452	.6122	1.5423				
rg (10)	.1226	1.0870	.6122	1.2166				
rg (15)	.1226	1.0460	.6122	1.1249				
rg (20)			.6122	1.0818				
rg (30)	.1226	1.0000	.6122	1.0402				
$V(\hat{Y}_{HT})$	.1076		.5063		.2176		.0697	
MSE WR	$.2459 \times 10^{15}$		$.3107 \times 10^{14}$		$.2077 \times 10^{13}$		$.4948 \times 10^{12}$	

Table 3A
Confidence Levels for Intervals Constructed With $\hat{Y}_{HT}$ $(\hat{Z}_{HT})$ and Several Estimators of $V(\hat{Y}_{HT})$ $(V(\hat{Z}_{HT}))$ for n = 30

	Universe I				Universe II			
	Y		Z		Y		Z	
Variance estimator	.90	.95	.90	.95	.90	.95	.90	.95
prg5	.8408	.8984	.8539	.9089	.7723	.8278	.7937	.8482
prg10	.8691	.9215	.8890	.9398	.8958	.8644	.8372	.8984
prg15	.8775	.9298	.8901	.9387	.8288	.8712	.8534	.9026
prg20	.9555	.9817	.9660	.9859	.8770	.9215	.9387	.9670
prg30	.8901	.9330	.9026	.9482	.8346	.8791	.8524	.9094
WR	.8901	.9330	.9026	.9482	.8346	.8791	.8524	.9094
WRA	.8885	.9288	.9000	.9456	.8319	.8775	.8482	.9047
RRS	.8885	.9298	.9000	.9461	.8330	.8775	.8513	.9063
CS	.8503	.9016	.8707	.9330	.8152	.8681	.8382	.8974
SP	.8461	.9042	.8607	.9319	.8220	.8733	.8450	.8995
SPW	.8429	.9011	.8586	.9309	.8178	.8728	.8429	.8979
$V(\hat{Y}_{HT})$ $(V(\hat{Z}_{HT}))$	.9319	.9607	.9115	.9529	.9236	.9602	.9052	.9482

Table 3B
Confidence Levels for Intervals Constructed With $\hat{Y}_{HT}$ $(\hat{Z}_{HT})$ and Several Estimators of $V(\hat{Y}_{HT})$ $(V(\hat{Z}_{HT}))$ for n = 60

	Universe I				Universe II			
	Y		Z		Y		Z	
Variance estimator	.90	.95	.90	.95	.90	.95	.90	.95
prg5	.8189	.8754	.8545	.9141	.7874	.8398	.8461	.8848
prg10	.8806	.9194	.8963	.9435	.8335	.8838	.8660	.9141
prg15	.8890	.9340	.8869	.9414	.8524	.9100	.8744	.9298
prg20	.8932	.9382	.9058	.9623	.8513	.9110	.8744	.9183
prg30	.9026	.9435	.9058	.9529	.8702	.9257	.8932	.9351
WR	.9110	.9487	.9100	.9602	.8691	.9309	.8932	.9298
WRA	.9026	.9476	.9068	.9550	.8649	.9309	.8901	.9278
RRS	.9068	.9476	.9079	.9571	.8681	.9309	.8911	.9278
CS	.8691	.9152	.8848	.9319	.8628	.9236	.8806	.9278
SP	.8586	.9194	.8848	.9351	.8754	.9225	.8869	.9319
SPW	.8524	.9131	.8817	.9298	.8639	.9194	.8838	.9246
$V(\hat{Y}_{HT})$ $(V(\hat{Z}_{HT}))$	.9215	.9529	.8995	.9508	.9215	.9508	.8995	.9550

Table 3C

Confidence Levels for Intervals Constructed With $\hat{Y}_{HT}$ ($\hat{Z}_{HT}$) and Several Estimators of $V(\hat{Y}_{HT})$ ($V(\hat{Z}_{HT})$) for n = 150

Variance estimator	Universe I				Universe II			
	Y		Z		Y		Z	
	.90	.95	.90	.95	.90	.95	.90	.95
prg5	.8508	.8822	.8272	.8979	.8272	.8874	.8115	.8560
prg10	.8874	.9189	.8796	.9162	.8246	.9058	.8534	.9110
prg15	.9031	.9372	.9136	.9476	.8325	.8953	.8456	.9084
prg20	.9136	.9555	.9189	.9555	.8639	.9136	.8639	.9215
prg30	.9136	.9529	.9241	.9581	.8351	.9162	.8456	.9058
WR	.9267	.9634	.9346	.9738	.8377	.9189	.8586	.9084
WRA	.8189	.9529	.9189	.9686	.8089	.8874	.8456	.8822
RRS	.9215	.9634	.9189	.9634	.8272	.9084	.8403	.9031
CS	.8848	.9476	.9031	.9607	.8429	.9005	.8508	.9005
SP	.8927	.9450	.9136	.9581	.8482	.9031	.8560	.9005
SPW	.8639	.9319	.8901	.9424	.8063	.8901	.8377	.8874
$V(\hat{Y}_{HT})$ ($V(\hat{Z}_{HT})$)	.9267	.9581	.9162	.9529	.9110	.9607	.8979	.9503

Table 3D

Confidence Levels for Intervals Constructed With $\hat{Y}_{HT}$ ($\hat{Z}_{HT}$) and Several Estimators of $V(\hat{Y}_{HT})$ ($V(\hat{Z}_{HT})$) for n = 300

Variance estimator	Universe I				Universe II			
	Y		Z		Y		Z	
	.90	.95	.90	.95	.90	.95	.90	.95
prg5	.7801	.8534	.8901	.9162	.7120	.7749	.8377	.9215
prg10	.8325	.9005	.9215	.9529	.7435	.8325	.8953	.9476
prg15	.8744	.9215	.9581	.9895	.7644	.8534	.8796	.9267
prg20	.8744	.9319	.9424	.9738	.8011	.8796	.9110	.9476
prg30	.8848	.9424	.9686	.9895	.7853	.8691	.9058	.9634
WR	.9162	.9529	.9791	.9895	.8011	.8744	.9100	.9581
WRA	.8796	.9319	.9581	.9843	.7382	.8325	.8691	.9267
RRS	.8796	.9424	.9581	.9895	.7539	.8325	.8796	.9424
CS	.8639	.9319	.9529	.9895	.7906	.8586	.8901	.9581
SP	.8744	.9215	.9581	.9895	.7906	.8639	.8901	.9581
SPW	.8272	.9058	.9372	.9943	.7278	.8272	.8586	.9372
$V(\hat{Y}_{HT})$ ($V(\hat{Z}_{HT})$)	.9162	.9634	.9162	.9529	.9058	.9581	.9058	.9424

Table 7

Universe Parameter Studies (Ratios to $V_{SYS}(\cdot)$)

		N	ρ	$-(\frac{1}{n-1})$	$V_{SYS}(\cdot)$	$V_{WR}(\cdot)$	VS	$V_{SRS}(\cdot)$	V_{SRS} [Ratio,X]	V_{SYS} [Ratio,X]	V_{WR} [Ratio,Z]	V_{SRS} [Ratio,Z]
Universe I	Y	30	-.00715	-.03448	$.14275 \times 10^{12}$	1.2620	1.2542	3.9321	1.4967	1.4262	1.4864	1.9820
		60	-.00345	-.01695	$.71766 \times 10^{11}$	1.2551	1.2395	3.8899	1.4805	1.3910	1.4783	1.9606
		150	-.00157	-.00671	$.27620 \times 10^{11}$	1.3045	1.2634	3.9779	1.5139	1.3055	1.5365	2.0049
		300	-.00030	-.00334	$.16424 \times 10^{11}$	1.0969	1.0275	3.2535	1.2382	1.0219	1.2919	1.6397
	Z	30	-.00409	-.03448	$.10758 \times 10^{08}$	1.1348	1.1243	7.2438	2.0244			
		60	-.00289	-.01695	$.50625 \times 10^{07}$	1.2059	1.1831	7.6555	2.1395			
		150	-.00073	-.00671	$.21760 \times 10^{07}$	1.1221	1.0688	7.0097	1.9588			
		300	-.00144	-.00334	$.69666 \times 10^{06}$	1.7524	1.5856	10.6490	2.9753			
Universe II	Y	30	.00143	-.03448	$.18764 \times 10^{12}$	.9601	.9542	2.9914	1.1386	1.0927	1.1308	1.5078
		60	.00057	-.01695	$.93099 \times 10^{11}$	.9675	.9555	2.9985	1.1413	.9685	1.1396	1.5114
		150	.00095	-.00671	$.41117 \times 10^{11}$	.8763	.8487	2.6721	1.0170	.8858	1.0321	1.3467
		300	.00159	-.00334	$.26584 \times 10^{11}$	.6777	.6348	2.0101	.7650	.7810	.7982	1.0130
	Z	30	.00139	-.03448	$.12700 \times 10^{08}$	.9613	.9524	6.1361	1.7148			
		60	-.00096	-.01695	$.57591 \times 10^{07}$	1.0600	1.0400	6.7295	1.8807			
		150	.00058	-.00671	$.26534 \times 10^{07}$	.9202	.8765	5.7485	1.6064			
		300	-.00019	-.00334	$.11505 \times 10^{07}$	1.0611	.9601	6.4482	1.8017			

VARIANCE ESTIMATION FOR STATE ESTIMATES FROM THE EXPANDED CURRENT POPULATION SURVEY

Lawrence Cahoon
U.S. Bureau of the Census

I. INTRODUCTION

The Comprehensive Employment and Training Act of 1973 (CETA) provides for the allocation of Federal funds to prime sponsors within the individual States on the basis of the "relative number of unemployed persons within the State as compared to such numbers in all States." [11]

At the request of the Bureau of Labor Statistics (BLS), the Census Bureau designed an expansion to the Current Population Survey (CPS) to produce State estimates that meet the reliability requirements of BLS. This required the selection of additional sample in approximately half the States.

This paper presents a general overview of the proposed variance estimation procedure for those States where an additional sample was chosen. These are two main areas of interest. These are the use of a collapsed stratum variance estimator and the use of a weighted average of sample data and census data variance estimators.

The usual procedure when a collapsed stratum variance estimator is used is to form each collapsed stratum from two of the original strata. Consideration is given to forming collapsed strata containing two or more of the original States. This procedure is evaluated from a mean-square error viewpoint.

The use of a census data variance estimate is considered due to the few numbers of sample areas in each State. While such a variance estimate is correct only at the time of the census, the estimate is being proposed in order to reduce the mean square error of the final variance estimate.

In order to facilitate the discussion of these two areas, a brief description of the CPS design and the design of the supplemental sample are given in Section II. A more complete description of the CPS design is given in [12]; a more complete description of the supplemental sample is given in [3]. Also discussed in Section II are some initial considerations in the variance estimation and aspects of the variance estimation which are preliminary to the two main issues discussed in this paper. The discussion of the collapsed stratum variance estimation, the methods for forming the collapsed strata, and the evaluation of these three methods are contained in Section III. The discussion of the two proposed census data variance estimators and the reasons for their use is given in Section IV.

II. INITIAL CONSIDERATIONS

A. The CPS Sample Design

Under the current CPS design the United States is divided into 1,924 primary sampling units (PSU's). These PSU's are grouped into 376 strata. One hundred and fifty-six of these strata contain only one PSU; the PSU's in these strata are included in the sample with certainty and are designated as self-representing (SR). The remaining PSU's are grouped in 220 strata with each stratum containing more than one PSU; the PSU's in these strata are designated as nonself-representing (NSR). The creation of these strata was done with the intention of obtaining the best national estimates and thus strata frequently cross State lines. In each stratum containing NSR PSU's a single PSU was selected with probability proportionate to size. Additionally, the 220 strata are grouped into 110 pairs. From each pair one stratum is selected at random (i.e., with equal probability). One PSU was chosen from the selected stratum with probability proportionate to size. Selection of the PSU's was independent for the two procedures.

B. The Sample Design in the States Where A Supplemental Sample was Chosen

In each of the States where additional sample was necessary in order to obtain the required degree of reliability on the State estimates, a supplemental sample, referred to as the CETA sample, was designed which attempts to maximize the use of the national CPS sample. Those PSU's which are self-representing in the CPS national design are retained as self-representing in the CETA State design.

The CETA design as it relates to the NSR PSU's is more complex. A requirement of the CETA design was that all NSR PSU's within a State be represented by a sample PSU within the State. The CPS strata cross State lines; therefore, at the first stage of the CETA design the in-State portion of each CPS national stratum was defined to be a CETA State stratum. These State strata were than divided into two groups. The first group contains those State strata which do not contain a sample PSU. The PSU's within these strata were regrouped into a new set of State strata. A single sample NSR PSU was selected within each stratum with probability proportionate to size. The second group of State strata were those which do contain sample PSU's. The State strata contained in this second group are retained and no additional sample PSU's are chosen within these State strata.

As a result of this procedure, a different selection of national CPS sample PSU's would have generated a different set of CETA State strata and a different set of CETA sample PSU's. Thus the strata definitions for the CETA design are random events. Nevertheless, the procedure was such that overall probabilities of selection were determinable and the resulting sample unbiased.

The estimation procedure for the CETA sample is similar to that used for the national CPS sample. A simple unbiased estimate is prepared by multiplying the value for each characteristic for each sample unit by the probability of selection of the sample unit. A noninterview adjustment by State is

made next to account for nonresponse. A first-stage ratio estimate is then produced by State, based on 1970 census totals, to adjust for differences in population characteristics in the sample PSU's and in the entire State. A national second-stage ratio adjustment is then made to the sum of the State first-stage ratio estimates based on the age, sex, race distribution of the United States population.

The discussion which follows focuses on the variance estimator of the unbiased estimate of population totals. The results presented may be extended to estimates produced at the suceeding stages of estimation without major modifications.

C. The Effects of the Random Strata Definitions on the Variance

The creation of the redefined strata was dependent upon the CPS strata in the States which are represented by national CPS sample PSU's in the in-State portion of each CPS stratum. This resulted in strata definitions for CETA in the supplemental States being random events.

We can express the variance over all possible samples for an estimate, Y, of the population total for a given characteristic as

$$\mathrm{Var}(Y) = E_1\,(\mathrm{Var}_2(Y)) + \mathrm{Var}_1(E_2(Y)).$$

The condition variance, $\mathrm{Var}_2(Y)$, and the conditional mean, $E_2(Y)$, are evaluated over all possible samples given a fixed strata definition. E_1 and Var_1 are evaluated over the range of possible strata definitions. We focus first on the term $\mathrm{Var}_1(E_2(Y))$. As was indicated in Section II-B, the expected value of the unbiased estimate over all possible samples given any fixed set of strata is a constant: i.e., $E_2(Y)$ is a constant. Thus $\mathrm{Var}_1(E_2(Y))$ is zero. $\mathrm{Var}_2(Y)$ is the variance of the sample estimate if the strata definitions were not random events; over all possible samples $\mathrm{Var}_2(Y)$ is an unbiased estimate of $E_1(\mathrm{Var}_2(Y))$. We propose to estimate $E_1(\mathrm{Var}_2(Y))$ in the usual fashion by $\mathrm{Var}_2(Y)$.

D. Variance Estimation in the Self-Representing PSU's

The only component of variance in the SR PSU's is the within-PSU variance. This variance will be estimated in the same manner as is done for the SR strata in the CPS design with a few minor modifications.

E. Variance Estimation in the Strata Containing NSR PSU's - An Introduction

The primary problem encountered in variance estimation for the estimates from the strata containing NSR PSU's is that there are a relatively few number of such strata in each supplemented State. This makes it difficult to obtain a variance estimate which can be regarded as reliable. The design in these strata which resulted from the CETA supplementation to the CPS meant that, with only a few exceptions, each stratum is represented by sample from a single PSU.

1. Estimation of the Within-PSU Variance in the NSR PSU's.

   The within-PSU component of variance for the NSR PSU will be estimated in exactly the same manner as for SR PSU's.

2. Estimation of the Total Variance for the Strata Containing NSR PSU's.

   The estimation of the total NSR variance is to be the weighted average of three variance estimates. The first estimate will be obtained by means of sample data using a collapsed stratum variance estimate. The second and third estimates will utilize the sample data estimate of within-PSU variance and two different estimates of between-PSU variance obtained using Census data. These three variance estimates are described in greater detail in the following sections.

III. Estimation of Total NSR Variance from Sample Data

A. Introduction and Theory

Hansen, Hurwitz and Madow [4] give the following formula for a collapsed stratum variance estimate:

$$\sum_{g}^{G} \frac{L_g}{L_g - 1} \sum_{h}^{L_g} \left(x'_{gh} - \frac{A_{gh}}{A_g}\, x'_g\right)^2 \qquad (1)$$

where A_{gh} is the population of stratum h in group g, x'_{gh} is the estimate for stratum h in group g, A_g is the population of group g, and x'_g is the estimate for group g.[1]

Normally L_g is taken as two if a collapsed stratum variance is to be used. The research discussed below was undertaken to determine the optimum size for L_g and whether the size of the groups can be varied to obtain "better" estimates of the variance for the sample which resulted from the CETA design. Methods are developed for approximating the variance of the estimator and the bias in the estimator. These approximations are then used to define groups into which the strata are placed. Three States are considered in detail. As a result of the research it was decided, with a few exceptions, to place all strata containing NSR PSU's in a single group for all supplemented States. As we are primarily interested in an accurate measure of the variance of the unemployment estimate, the evaluations utilize unemployment data.

It can be shown that the collapsed stratum variance estimate, formula (1), tends to be an over-estimate of the true variance. Hansen, Hurwitz and Madow [4] show that formula (1) is a biased estimate; specifically the expected value of formula (1) is

$$\sum_{g}^{G}\sum_{h}^{L_g}\sigma^2_{x'_{gh}} + \sum_{g}^{G}\left\{\frac{1}{L_g - 1}\left[V^2_{A_{g(h)}} - 2V_{A_{g(h)}}\sigma^2_{x'_{g(h)}}\right]\sum_{h}^{L_g}\sigma^2_{x'_{gh}}\right\}$$

$$+ \sum_{g}^{G}\frac{1}{L_g - 1}\sum_{g}^{L_g}\left(X_{gh} - \frac{A_{gh}}{A_g}X_g\right)^2 . \qquad (2)$$

where

$$V_{A_{g(h)},\sigma^2_{x'_{g(h)}}} = \frac{\sum_{h}^{L_g} A_{gh}\sigma^2_{x'_{gh}}}{\bar{A}_g\,\sigma^2_{x'_g}} - 1 , \qquad (3)$$

with

$$\sigma^2_{x'_g} = \sum_{h}^{L_g}\sigma^2_{x'_{gh}} , \quad \bar{A}_g = \frac{1}{L_g}\sum_{h}^{L_g} A_{gh} \qquad (4)$$

and where

$$V^2_{A_{g(h)}} = \frac{\sum_{h}^{L_g} A^2_{gh}}{L_g\bar{A}^2_g} - 1. \qquad (5)$$

The second and third terms in expression (2) are the biases in the collapsed stratum variance estimate. Expression (5) can be recognized as the relvariance of the strata sizes. We wish to approximate the bias in the estimated variance depending on the size of the groups and the composition of the groups.

In order to simplify the evaluation of the bias in the variance estimates, assume that the total variance for a stratum is proportional to the size of the stratum. That is, assume

$$\sigma^2_{x'_{gh}} = cA_{gh}.$$

This is a good assumption for national CPS and for the supplemented States in CETA where the between-PSU variance is a small proportion of the total variance. Under this assumption equation (3) becomes

$$V_{A_{g(h)},\sigma^2_{x'_{g(h)}}} = \frac{\sum_{h}^{L_g} A^2_{gh}}{L_g\bar{A}^2_g} - 1 = V^2_{A_{g(h)}}$$

Thus the second term in expression (2) becomes

$$-\sum_{g}^{G}\frac{1}{L_g - 1}V^2_{A_{g(h)}}\sum_{h}^{L_g}\sigma^2_{x'_{g(h)}} .$$

Thus the bias for group g from the second term in expression (2) expressed as a percent of the variance for group g is

$$\frac{100}{L_g - 1}\left[1 - \frac{\sum_{h=1}^{L_g} A^2_{gh}}{L_g\bar{A}^2_{gh}}\right] \text{ percent.} \qquad (6)$$

Since this term is dependent only on the size of the strata we may properly designate it as the bias due to differences in stratum sizes within group g. Unless the strata vary widely in size this term tends to be small.

The third term in expression (2) is the bias due to differences in the characteristics of the strata. This term can be determined directly from census values; however, those calculations are correct only for one point in time. We will assume, with some caution, that those calculations will be indicative of the present magnitude of this term. While we can approximate the magnitude of this term from census data, to obtain an estimate of the relative bias due to this term we must know the total NSR variance of the estimate of level. We can approximate this value.

The total variance for a group of strata for unemployment items can be approximated by $b_s X_g$ where $X_g = \sum_h X_{gh}$ is the census value for the item for group g and b_s is the product of the State NSR design effect and the State NSR sampling interval.[2] This is an acceptable approximation to the NSR variance as the number of unemployed persons is a small percentage of the total population. The computation of the design effects for States is based on the work by C. Dippo using 1960 census data with one change. Based on data obtained since the CETA supplementation we have assumed a 1.1 within-PSU design effect instead of the 1.4 within-PSU effect used by C. Dippo [3].[3]

The approximate percent bias for each group due to differences in the characteristic across strata can then be approximated by:

$$\frac{100\ L_g}{L_g - 1}\sum_{h=1}^{L_g}\left(X_{gh} - \frac{A_{gh}}{A_g}X_g\right)^2 \Big/ b_s X_g \qquad (7)$$

The numerator of expression (7) is the bias within group g and is obtained from the third term of expression (2). The denominator is the approximate variance in the State for group g based on the above design effect.

B. Three Methods of Forming Collapsed Strata

For the convenience of our discussion we will describe three methods for grouping the strata within the State and associate an estimate with each of these methods. The three methods are:

Method I. Place all strata in a single group.

Method II. For this method all strata are placed in groups of size two (i.e., $L_g = 2\ V_g$); if there is an odd number of strata, one group consists of three strata. The pairing of strata for the group is done so as to have strata with similar size and characteristics in the same group. This is done so as to minimize the bias in the variance estimate.

Method III. For this method the strata are placed in groups of varying size, the only constraint being that all strata in a given group be of similar size and have similar characteristics. Methods I and II are special cases of this method of grouping.

C. Comparison of the Bias for the Three Methods

We will consider three States as examples in the computation of the biases. Subsequently, we will compute an approximate mean square error for the estimates discussed here. The characteristic of interest in these evaluations is the 1960 census unemployment level.

1. Arkansas. The population, the 1960 census unemployment rates, and the projected unemployment rates for the strata in Arkansas are given in table 1. The projected unemployment rate was obtained from the step-wise regression program used to determine the strata definitions for the supplemented States in the CETA expansion [3].

Table 1 ARKANSAS
Stratum Unemployment Characteristics

Stratum	In-State 1970 Population	Projected 1970 Unemployment Rate	1960 Unemployment Level
577	131501	0.0183	1555
590	88882	0.0282	2494
757	158725	0.0227	2656
658	76859	0.0216	1402
681	124112	0.0176	2006
914	87124	0.0190	1881
945	149436	0.0222	2617
AR1	110080	0.0154	1924
AR2	126966	0.0209	2582
AR3	112639	0.0226	2847
AR4	122101	0.0244	2571
AR5	125169	0.0304	2859
Total	1413597	--	27394

The tabulations in table 1 were used to form the groups and evaluate the biases for each of the three methods. For Arkansas, differences in stratum population were not considered in forming the group since the bias resulting from these differences is small. To obtain an estimate of the bias for Method I, all strata were placed in a single group. For Method II, a more complicated procedure was used. First, it was felt that the groups should be formed based upon the 1970 projected unemployment rates since these rates were used to form the strata in the State and, second, we wished to have strata with similar characteristics in the same group. Therefore, the groups were formed according to the following procedure. The two strata with the lowest projected unemployment rate formed the first group. The next group contained the two strata with the lowest projected unemployment rates among the strata remaining. This procedure was continued until six groups of two strata each were formed. This procedure results in the minimum bias possible among all possible groupings with two strata per group when the bias is computed based on the 1970 projected unemployment rates.

In forming the groups for Method III we used the same considerations as were used for Method II. Again we wished to have strata with similar characteristics in the same group. We again used the projected unemployment rates to form the groups. Two constraints were imposed in forming the groups. First, at least one group had to contain more than two strata; this prevented Methods II and III from resulting in the same set of groups. The second constraint was that the projected unemployment rate for all strata in group i be less than the projected unemployment rate for each stratum in group j if $j > i$. These constraints allow for several different groupings of the strata in the State. Each of these were considered as a possible grouping for the strata. Once these constraints have been satisfied there are several possible sets of groups for Method III. It was decided to choose the grouping which satisfied the given constraints and which minimized the relative mean square error of the variance estimate when the bias is computed based on the 1960 census unemployment data. This methodolgy actually gives an unfair advantage to Method III because the characteristic of interest is used to determine the stratification. The methodology used to estimate the relative mean square error of the variance estimate is described below. For each of the groupings resulting from the three methods formula (2) can be used to estimate the total NSR variance. For the State of Arkansas the groups resulting from each of these three methods are:

Method I. A single group of all twelve NSR strata.

Method II. Six groups -- (AR1,681), (577,914), (AR2,658), (945,AR3), (657,AR4), (590,AR5).

Method III. Three groups -- (577,681,AR1), (914,658,AR2,945,657,AR3,AR4), (590,AR5).

The estimates of the relative bias in the variance estimates when each of these three sets of groups are used to estimate the variance are given in table 2.

Table 2 ARKANSAS
Bias in the Estimates of Variance

	Method I.	Method II.	Method III.
Degrees of Freedom	11	6	9
Bias Due to Differences in Stratum Population	-0.36%	-2.74%	-1.16%
Bias Due to Differences in Stratum Characteristics	6.70%	6.12%	3.64%
Net Bias	6.34%	3.38%	2.48%

The interesting result from Table 2 is that the use of the collapsed stratum variance estimation procedure does not minimize the bias when each group contains only two strata.

2. South Dakota. The data used to form the groups for the Strata in South Dakota for each of the three methods is given in table 3. The resulting biases for the three methods are given in table 4.

Table 3 SOUTH DAKOTA

Stratum	In-State 1970 Population	Projected 1970 Unemployment Rate	1960 Unemployment Level
464	48336	0.0087	813
SD1	31269	0.0069	440
SD2	34485	0.0091	441
SD3	31748	0.0098	554
SD4	30930	0.0099	505
SD5	36587	0.0104	539
SD6	40338	0.0115	699
SD7	32348	0.0119	377
SD8	37718	0.0126	348
SD9	30965	0.0129	693
SD10	31632	0.0157	621
SD11	30277	0.0292	748
TOTAL	415633	--	6778

Table 4 SOUTH DAKOTA
Bias in the Estimates of Variance

	Method I.	Method II.	Method III.
Degrees of Freedom	11	6	10
Bias Due to Differences in Stratum Population	-0.20%	-1.33%	-0.20%
Bias Due to Differences in Stratum Characteristics	5.21%	6.84%	2.67%
Net Bias	5.01%	5.51%	2.47%

The groups for South Dakota are:

Method I. A single group of all 12 NSR strata.

Method II. Six groups of two strata each -- (SD1, 464), (SD2,SD3), (SD4,SD5),(SD6,SD7), (SD8,SD9), (SD10,SD11).

Method III. Two groups -- (SD1,464,SD2,SD3,SD4, SD5,SD6,SD7,SD8),(SD9,SD10,SD11)

3. Idaho. For Idaho the bias in the variance estimate is extremely large regardless of which of the three methods is used. This illustrates the need to utilize a census data variance estimator as a part of the total variance estimate. The results of the bias calculation are given in table 6. The tabulations used to form the groups and to compute the approximate bias are given in table 5. The groupings for each of the three methods are:

Method I. One group containing all six strata.

Method II. The groups are (ID1,840), (ID2,807), (ID3,ID4).

Method III. ID1,840,ID2), (807,ID3,ID4).

Table 5 IDAHO

Stratum	In-State 1970 Population	Projected 1970 Unemployment Rate	1960 Unemployment Level
807	48205	0.0237	856
840	55151	0.0099	832
ID1	49129	0.0094	837
ID2	50089	0.0143	603
ID3	47983	0.0242	790
ID4	53943	0.0494	2329
TOTAL	304499	--	6247

Table 6 IDAHO
Bias in the Estimates of Variance

	Method I.	Method II.	Method III.
Degrees of Freedom	5	3	4
Bias Due to Differences in Stratum Population	-0.06%	-0.27%	-0.29%
Bias Due to Differences in Stratum Characteristics	53.31%	46.37%	43.93%
Net Bias	53.25%	46.10%	43.64%

The large biases observed in table 6 in the variance estimators for Idaho are due to the large differences between stratum ID4 and the other strata in the State. Stratum ID4 has twice the unemployment rate of any other NSR stratum in the State; this difference remained unchanged between 1960 and 1970. This example illustrates that we should not blindly use the collapsed stratum variance estimation technique; rather, we should do a careful evaluation of the procedure on a State-by-State basis.

D. The Variance of the Variance Estimates

We wish to evaluate the methods based on the mean square error of the variance estimate. First we must approximate the the variance of the variance estimate.

Hansen, Hurwitz and Madow [4] give the following formula for the relvariance of the estimated variance when proportionate stratified sampling is used and when the strata are of equal size and the within stratum variances are all the same.

$$Z^2 = \frac{1}{n}\left(\bar{\beta} - \frac{\bar{n}-3}{\bar{n}-1}\right), \qquad \bar{\beta} = \frac{\sum_{h=1}^{L} \mu_{4h}}{L\bar{s}^4}$$

where μ_{4h} is the fourth central moment of stratum h, L is the number of strata, $\bar{s}^2$ is the within stratum variance, $\bar{n}$ is the number PSU's selected per stratum and n is the total number of PSU's selected. The assumption made in this formula are restrictive and are not met exactly in the CETA sample. However, the assumptions are not so restrictive that formula (8) cannot be used as an approximation to the relvariance of the estimate of variance. In using formula (8) we will make the additional assumption, that μ_{4h} is the same for all strata. Thus the expression for $\bar{\beta}$ reduces to $\bar{\beta} = \mu_4/\bar{s}^4$. Thus $\bar{\beta}$ is the kurtosis of the within stratum distribution for the characteristic. Since we are using a sample where the variates take on the values 1 or 0 (i.e., 1 if unemployed and 0 if

employed) we have a binomal distribution and

$$\bar{\beta} = \frac{1}{PQ} - 3 \qquad (9)$$

where P is the proportion of people with the characteristic. For Method I the relvariance of the variance estimate is

$$Z_2^2 = \frac{1}{n}\left(\bar{\beta} - \frac{n-3}{n-1}\right) \qquad (10)$$

For Method II, where $\bar{n} = 2$, the relvariance is

$$Z_1^2 = \frac{1}{n}(\bar{\beta} + 1) \qquad (11)$$

The relvariance of the estimate from Method III cannot be defined from formula (8). For this reason, we approximate its relvariance for Method III, Z_3^2, by linear interpolation based on the degrees of freedom for each method.

These approximations can than be used to approximate the relative mean square error, Rel-MSE, of the estimates.

We are primarily interested in obtaining accurate estimates of the variance of yearly averages. We do know that for unemployment items the variance of a yearly estimated average is approximately 20 percent of the variance of a monthly estimate. Based on this, we assume the same relationship for the variance of the variance estimate. Large variations from the factor of five rarely influences the choice of methods.

E. Comparisons of the Mean Square Error of the Three Methods

Utilizing this assumption and the theory previously developed, tables 7 and 8 present the Rel-MSE for each of the three methods for Arkansas and South Dakota respectively.

Since the major concern is with estimating State unemployment and variance of that estimate values of P, the percent of the population unemployed, between 0.03 and 0.05 are of primary interest. The Rel-MSE for these values of P are given in Tables 7 and 8. It can be shown from equations (12) and (13) that which method minimizes the Rel-MSE is not dependent upon the value of P.

Table 7 ARKANSAS
Relative Mean Squared Error of the Estimates of Variance

P = 0.03	Rel-MSE Monthly Estimate	Rel-MSE Yearly Average	Degrees of Freedom
Method I	2.5492	0.5131	11
Method II	2.6978	0.5405	6
Method III	2.6064	0.5218	9
P = 0.05			
Method I	1.4400	0.2912	11
Method II	1.5886	0.3186	6
Method III	1.4972	0.2999	9

Table 8 SOUTH DAKOTA
Relative Mean Squared Error of the Estimates of Variance

P = 0.03	Rel-MSE Monthly Estimate	Rel-MSE Yearly Average	Degrees of Freedom
Method I	2.5477	0.5116	11
Method II	2.6997	0.5424	6
Method III	2.5761	0.5157	10

Table 8 (cont'd)

P = 0.05	Rel-MSE Monthly Estimate	Rel-MSE Yearly Average	Degrees of Freedom
Method I	1.4385	0.2897	11
Method II	1.5705	0.3205	6
Method III	1.4669	0.2939	10

The calculations of the previous two sections are relative to only unemployment and do not take into account other characteristics. Except for some minor considerations used in Method III the calulations do not allow for changes in the characteristics of the strata over time. It is felt that changes in the characteristics of the strata over time should in general affect Method II the most and Method I the least. On the basis of minimum relative mean square error. Based on this criterion, we choose Method I for Arkansas and North Dakota. The magnitude of the bias in the estimates for Idaho indicated that special consideration should be given to that State. In most of the remaining supplemented States a similar analysis indicates that Method I is to be preferred.

The actual method chosen does not make a large difference in the relative-mean-square error of the variance estimator. The frequently used procedure is to form groups of size two, i.e., use Method II. This data indicates that while there is little difference among the methods, Method II is the worst of the three.

IV. Estimation of Total NSR Variance from Census Data

The estimation of the total NSR variance from census data, in fact, utilizes the census data to estimate only the between-PSU component of variance. The within-PSU component of variance will be estimated from sample data as outlined in Section II-D.

The first census data estimate of between-PSU variance takes the usual form of the variance over all possible samples. This variance estimator is:

$$C^2 \left(\sum_h \sum_i \frac{P_h}{P_{hi}} X_{hi}^2 - \sum_h X_h^2 \right)$$

where

P_h is the 1970 census population in stratum h, P_{hi} is the 1970 census population in PSU i in stratum h, X_h is the 1960 census total for the characteristic for stratum h, X_{hi} is the 1960 census total for the characteristic for PSU i in stratum h, $C = \hat{u}/(\sum_h X_h)$, and $\hat{u}$ is the current survey estimate for the characteristic.

The term C^2 is included in expression (8) to adjust the variance estimate for the differences in the level of the estimate between 1960 and the time of the survey.

The second census data between-PSU variance estimate is not the typical variance estimate. Instead, it is a direct measure of the squared error due to the given selection of sample PSU's within the NSR strata. The use of this variance estimator was suggested by Gary Shapiro of the Bureau of the Census. The variance estimate is

$$C^2 \left[\sum_h \left(\frac{P_h}{P_{hi}} X_{hi} - X_h \right) \right]^2$$

Here i is the PSU in stratum h which is in sample. This formula provides the best measure of the actual squared error resulting from the selection of the NSR PUS's. It accounts for the difference between the characteristics of the PSU's in sample and the characteristics of the State taken as a whole. Thus this form of variance estimator is specific to the actual set of PSU's selected whereas formula (8) is not. Formula (8) provides a variance estimate over all possible samples.

Current plans are that the final variance estimate be a weighted average of the sample data and the two census data variance estimates. The within-PSU variance is to be estimated entirely from census data. The census data variance estimate will be used to reduce the mean square error of the between-PSU variance estimate. The weights to be used have not been determined.

REFERENCES

[1] Cochran, W.G., "Sampling Techniques," John Wiley and Sons, New York (1963).

[2] Dippo, C., "CPS-CETA Documentation: Strata Listing for States Supplemental for CETA," Internal Census Bureau Document, February 25, 1976.

[3] Dippo,C., "Expansion of CPS to Provide Reliable State Estimates of Unemployment," Proceedings of the American Statistical Association, August, 1975.

[4] Hansen, M.H., Hurwitz, W.N., and Madow, W.G., "Sampling Survey Methods and Theory," Vol. I, Vol. II, John Wiley and Sons, New York (1953).

[5] Keyfitz, Nathan, "Estimates of Sampling Variance When Two Units are Selected from Each Stratum," Journal of the American Statistical Association, 52 (1957) pp. 503-510.

[6] McCarthy, P.J., "Pseudo-Replication: Half Samples," Review of the International Statistical Institute 37; Number 3 (1969) pp. 239-264.

[7] Shapiro, Gary M., "Keyfitz Method of Estimating Variance and Its Application to the Current Population Survey," Internal Census Bureau Document, September 1, 1966.

[8] Thompson, M.M., Shapiro, G., "The Current Population Survey: An Overview," Annals of Economic and Social Measurement, Vol. 2, No. 2 (1973).

[9] Tepping, B.J., "Variance Estimation in Complex Surveys," Proceedings of the American Statistical Association, August 1968.

[10] Woodruff, R.S., "Simple Method of Approximating the Variance of a Complicated Estimate," Journal of the American Statistical Association, June 1971, pp. 411-414.

[11] _____, "Comprehensive Employment and Training Act of1973," Public Law 93-203, 93rd Congress, S-1559, December 28, 1973.

[12] _____, "U.S. Bureau of the Census, "The Current Population Survey--A Report on Methodology," Technical Paper Number 7, 1963, (This paper is currently being revised).

[1] A_{gh} and A_g need not be population totals but may be the value of any known characteristic correlated with Ex'_{gh} and Ex'_g.

[2] This is a slight deviation from our earlier assumption that stratum variance is proportional to the size of the characteristic. The assumptions are approximately equivalent.

[3] Based on more recent data the within-PSU des design effect is being revised. Current indications are that the design effect is between 1.3 and 1.5.

THE ANALYSIS OF SAMPLING ERRORS

K.P. Krotki, Statistics Canada; L. Kish, The University of Michigan;
R.M. Groves, The University of Michigan

1. Introduction

This investigation is based on eight fertility surveys from five countries (South Korea, Taiwan, Malaysia, Peru and the United States), all of them conducted before 1974. The unique aspect of this investigation is the large number and variety of sampling error results that are calculated and analyzed. We suggest methods for the analysis and presentation of sampling errors for future surveys. Continued work in this field will hopefully lead to a type of data bank containing sampling errors for a large number of statistics originating from a vareity of sample designs.

2. Methodology

2.1 Formulas and calculations of deft and roh values

Deft (the square root of deff, the design effect) and roh (the synthetic intra-class correlation coefficient) are presented for approximately 40 means on the total sample and on 24 subgroups from each survey. We will refer to these means as "characteristics" and the subgroups as "subclasses." The choice of these characteristics was a subjective process guided by a desire to achieve a wide variety of substantive issues and some variation in the sensitivity of the statistic to clustering effects.

The formulas used, in their most basic form, are:

$deft^2 = var(r) / (s^2/n)$ where r is the ratio mean for a characteristic, var(r) is the computed sampling variance, and s^2/n is the simple random sample variance (estimatable by (pq)/n in the case of a proportion p).

$roh = (deft^2 - 1) / (\bar{b} - 1)$ where $\bar{b}$ is the average cluster size measured as the sample size, n, divided by the number of clusters, a.

The sample mean, r, a ratio mean, is of the form (y/x) where, because of clustering, x (as well as y) is a random variable because of variation in cluster size. In order to calculate the variance or r we use the approximate formula:

$$var(r) \doteq (1/x^2)\left[var(y) + r^2 var(x) - 2r\,cov(x,y)\right].$$

Stratification and clustering are introduced into the calculation of var (r) in the standard fashion. The paired difference calculation was deemed appropriate in all the surveys. The samples on which the surveys were based were stratified, clustered areal probability samples. The sampling elements were women of child-bearing ages, and the primary sampling units (PSU's or clusters) were geographical units (e.g., counties, townships, city blocks).

Sampling errors were calculated for means and proportions of both the total sample, subclasses, and differences between subclass means. These consisted of differences (y/x - y'/x') for the same characteristic in two categories of the same variable; the computations of these variances contain two variances and a covariance term. To compute a "synthetic roh," the value of $\bar{b}$ for the difference of means uses the harmonic mean of the sample sizes for the two subclasses.

2.2 Portability

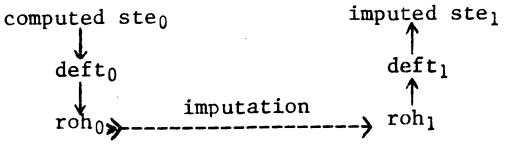
Our goal is to compute and present estimates of design parameters that can be used both simply and generally for diverse multipurpose designs. We think that portable estimates conveys the meaning we need. Portability refers to properties of the estimate that facilitate its use far from its source.

To illustrate, let us begin with the standard error, $ste(\bar{y})$, one computes for making inferential statements like $\bar{y} \pm t_p\, ste(\bar{y})$. Standard errors computed for one statistic can be imputed directly only to essentially similar survey designs. They are specific to the estimate $\bar{y}$ and depend on: a) the nature of the variables, b) their units of measurement, c) the nature and design of statistics derived from variables, d) sizes of the sample bases, which can vary greatly for subclasses, e) sizes of selections from sample clusters, f) nature and size of sampling units.

Design effects are considerably more portable than standard errors. They are widely used to modify simple random estimates $ste_{srs}(\bar{y})$ to guess at some ste(r) as $\left[deft \times ste_{srs}(\bar{y})\right]$. When we compute $deft = ste(r)/ste_{srs}(\bar{y})$, we remove the effects of the units of measurement and of the sample's aggregate size.

However, design effects for most subclasses diminish along with sample size, and using values of deft computed from the entire sample grossly exaggerates the actual effect of the design on subclasses. Also, deft values depend heavily on the sizes of sample clusters used.

We need portability to make inferences from one set of results to a set of variates with different values of $\bar{b}$. Values of roh are more portable for this purpose than deft or ste. We found usable stable relationships of roh for subclass means to roh for total sample means-much more stable than for values of deft or ste. Also we found relative stability of roh values across diverse subclasses for each characteristic from a sample; and similarities for similar characteristics across samples. Thus we propose the following indirect method of imputation from a computed standard error (ste_0) to an unknown one (ste_1):



We must, however, remain aware of factors that interfere with complete portability. The computed values of roh are also functions of the kind of sampling units used and of the selection procedures in several stages.

2.3 The use of roh and deft for imputation

We need to impute roh for subclasses from values computed for the entire sample or for similar type subclasses. Thus we need stability (portability) for roh values and we seem to find that for crossclasses. This type seems to cover most subclasses used in survey analysis. Crossclasses is a term we coined for subclasses that cut across clusters and strata used in the selection process. The sizes of sample clusters for each subclass are roughly $\bar{b}_s = \bar{b} M_s$, where M_s is the proportion of the subclass in the sample and $\bar{b}$ is for the entire sample. Design effects tend to decrease linearly almost to 1 as the crossclass size decreases and roh remains relatively constant. We must first impute some value $roh_1 = \lambda_1 roh_0$ from computed values of roh_0 and a correction factor λ_1. Then we estimate the unknown $deft_1$ from $deft_1^2 = 1 + \lambda_1 roh_0 (\bar{b}_1 - 1)$. We computed values of roh_0 based on means for the entire sample for each of 40 characteristics on each survey. We then computed and found values near (and slightly over) $\lambda_1 = 1$ for the diverse subclasses.

2.4 Summarizing sampling error results

Sampling errors computed from survey samples are themselves usually subject to great sampling variability. Many samples are not based on a large enough number of PSU's to yield sufficient precision for individual estimates for sampling errors. In addition, most surveys are highly multipurpose in nature and we must combine results from diverse statistics for joint decisions and designs. Some form for combining them must be sought, because combining their results is preferable to its alternatives. We argue against following the common practice of choosing a single variable among many for making inferences about the design and planning future designs.

Several methods were applied to the sampling error results in this investigation in order to identify underlying trends and relationships. Much of what was done was on an ad hoc basis as each survey presented its own idiosyncracies. Thus the methods shown here should be viewed more as a progress report than as final optimal techniques. Hopefully we have pointed out some approaches that may be applicable on a more general scale.

First, characteristics were listed by order of magnitude of roh. Another approach to arrive at the same information is to group supposedly "similar" characteristics and to calculate the average roh for each group. The mean and range of roh values for the characteristics within each group can serve as summary statistics. Measurements on the same characteristics at different points in time or under different survey conditions provide further data on the sampling behavior of these characteristics.

The study of sampling errors for subclasses is an important need because much survey analysis involves comparisons of subclasses. It is difficult to give guides for how the choice of subclasses should be made, but using measures which are candidates for independent variables in analysis of the data may be desirable. In this view, the characteristics would be analogous to the dependent variables. Comparison of sampling errors for the total sample and for the subclasses can give the survey designer an idea of how to impute in general from total results to subclasses. This is a common requirement since sampling errors cannot be calculated for all possible subclasses for each characteristic.

3. Empirical Results

The above described methodology was applied to the sampling errors calculated for eight fertility surveys in five countries. In this section we discuss in detail the results for one of these surveys.

Detailed analysis of sampling error results for Taiwan: General Fertility Survey (1973 KAP-4)

3.1 Sample design

The universe of 331 townships was divided into 27 strata using level of urbanization, education, and fertility. Within strata, townships were geographically ordered and 56 were selected systematically. Within selected townships the sample had three stages, yielding 5588 married women aged 20-39. The coefficient of variation of size among the 56 ultimate clusters is 0.03 for the entire sample; within the 24 subclasses used it ranges from 0.02 to 0.08.

3.2 Results for the total sample

Results for 40 characteristics are presented in Table 1. The characteristics are ordered from highest to lowest values of roh. Deft values follow this trend closely with minor exceptions due to slight differences in sample bases (n), hence cluster size (n/a). Note the large range of roh values (col. 4) for the 40 characteristics, essentially from 0 to 0.3. The quartiles are about 0.075, 0.025 and 0.015. These correspond to deff values of about 8.4, 3.2, and

Table 1

Taiwan Fertility Study (KAP), 1970, Means, Ste's, Deft's and Roh's for 40 Characteristics Together with Summary Roh Values for Subclasses and Differences[a]

Char. Type	Characteristic	1 Mean	2 Std. Error	3 Total Sample Deft	4 Total Sample roh	5 Sub-Class roh	6 (5)/(4)	7 Ave. rohd (c-r')
3	Sex preference scale	5.23	.053	5.41	.290	.334	1.15	.012
4	Approve contraception strongly	0.38	.034	5.28	.273	.350	1.28	.020
4	Approve sterilization	0.72	.029	4.75	.219	.251	1.15	.007
4	Should have many children	.037	.029	4.49	.194	.241	1.24	.015
4	Ideal first birth interval	20.86	.478	3.82	.140	.181	1.29	.006
3	Number preference scale	4.70	.053	3.59	.122	.186	1.52	.016
3	Husbands not wanted last pregnancy	0.24	.019	3.39	.106	.125	1.18	.010
4	Ideal marriage age	23.10	.076	3.23	.096	.115	1.19	.012
4	Expect sterilization	0.33	.020	2.98	.088	.107	1.22	.003
4	Approve abortion	0.24	.017	2.94	.078	.134	1.72	.014
2	Visited Health Station	0.47	.019	2.88	.074	.105	1.42	.009
4	Others should have < 3 children	0.66	.018	2.87	.074	.088	1.19	.007
3	Desired children < expected	0.06	.008	2.50	.057	.079	1.39	.002
2	Contraception from private MD	0.47	.018	1.96	.055	.090	1.63	.018
3	Ideal number of children	1.37	.018	2.42	.051	.063	1.23	.006
3	Husband's ideal number of children	3.24	.028	2.26	.048	.075	1.55	.014
2	Visited by health worker	0.37	.015	2.37	.047	.072	1.55	.005
3	Ideal number of boys	1.89	.014	2.08	.036	.043	1.22	.005
2	Plan no future contraception	0.10	.008	1.92	.028	.042	1.47	.007
6	Age at marriage	20.31	.072	1.86	.025	.041	1.62	.008
3	Wife-husband want same number of children	0.19	.010	1.83	.024	.037	1.55	.006
1	Able to have children	0.86	.008	1.81	.023	.028	1.22	.003
3	Desired number of children	3.54	.031	1.79	.023	.038	1.68	.005
2	Contraception started after pregnancy number	3.57	.042	1.55	.022	.040	1.86	.006
1	Husband's mother's number children	6.05	.059	1.72	.021	.036	1.74	.005
3	Expected total births	3.58	.030	1.68	.020	.040	2.06	.006
5	Literate wife	0.75	.010	1.67	.018	.042	2.31	.008
1	Number of live births	3.20	.037	1.65	.017	.032	1.86	.008
1	Wife's mother's number children	6.45	.051	1.62	.016	.020	1.25	.004
2	Ever used contraception	0.67	.010	1.61	.016	.020	1.28	.001
3	Want no more children	0.67	.010	1.56	.014	.014	1.01	-.003
1	First birth interval	15.14	.236	1.49	.013	.017	1.29	-.002
1	Open birth interval	45.22	.836	1.52	.013	.025	1.93	.003
5	Literate husband	0.92	.005	1.50	.013	.024	1.89	.007
2	Contraception before 1st pregnancy	0.02	.003	1.35	.011	.006	.050	.000
2	Currently using contraception	0.45	.010	1.45	.011	.006	0.57	-.002
1	Living sons number	1.54	.021	1.43	.011	.012	1.08	.002
1	Living children number	3.04	.029	1.39	.010	.017	1.75	.006
1	Pregnant now	0.12	.005	1.21	.005	.005	1.11	-.001
2	Induced abortions number	0.31	.012	1.19	.004	.012	2.72	.004
	Averages				.0592	.0790	1.436	.00652
	Ratios of means of col. 5/col. 4 and col. 7/col. 5					1.334		.083

[a]The characteristic type denotes: 1) fertility experience, 2) contraceptive practice, 3) birth preferences and desires, 4) attitudes, 5) socio-economic background, 6) demographic background.

2.5; these large factors arise because of the large number of elements, almost 100, per cluster. The mean roh on the total sample is 0.0592.

It is useful to observe the clear differences in roh values between the 6 classes of characteristics. Attitudinal variables are all in the first quartile, with roh value over 0.075. Birth preferences and desires are mostly in the top two quartiles, with roh values over 0.025. Contraceptive practice is spread evenly between the second quartile (0.075 - 0.025) and the second half under 0.025. Fertility experience variables are all in the lower half with roh values under 0.025. They are evenly spread among socio-economic (which, in this survey, only indicates literacy) and demographic variables. These three classes of variables (codes 1, 5 and 6) are contained in the lower half, with roh values under 0.025, while classes 3 and 4 are above that.

If roh values were unusually high for all variables, we should look either into causes for unusual segregation in the population or into the choice of small and homogeneous sampling units. However, roh's for demographic variables are not high. Their spread under 0.025 is similar to values found in other populations. Two explanations are possible for the high roh values for the subjective variables of attitides and birth preferences and desires. First, is is sociologically reasonable to think that when attitudes change rapidly, the spread of the change takes place unevenly and is clustered in areas. Second, clustering of the measured values can be caused by interviewer effects which are not separable from the effects of clusters themselves.

3.3 Results for subclasses

Clustering of values for subgroups of the sample was investigated for the 24 subclasses in Table 2 for each of the 40 characteristics. This vast amount of data is summarized in Column 5 of Table 1. Each entry is the mean of the rohs over the same 24 subclasses of Table 2. This mean subclass roh is shown as the ratio to the roh for the total sample (col. 6). Note that the mean subclass roh values parallel closely the total roh values. The ratios of subclass/total roh values do not vary greatly around their mean of 1.436. A more useful average is .0790/.0592 = 1.334, the ratio of the two mean values. This gives greater weight to the larger roh's where more fluctuations can be observed. A quick rule of thumb would guide the researcher to use the total roh times 1.33 to obtain subclass roh's. This yields

$$deff_{subclass} = [1 + 1.33\, roh_{Total}\, (b_{subclass} - 1)]$$

Column 4 of Table 2 presents values of roh for each subclass averaged over all 40 characteristics. Column 5 notes the ratios of these averages to the mean roh value of 0.0592 when the total sample is the base. For these values of subclass bases there exists no clear separation between socio-economic and demographic subclasses that we found for them as characteristics. Though the former tend to be a little higher, most of the variation is within the groups. The

Table 2

Taiwan Fertility Study (KAP), 1970, Deff's and Roh's for Twenty-four Subclass Variables Treated as Characteristics and Subclass Base.

		1	2	3	4	5	6	7
		Population Base			Subclass base		Differences	
		Prop.	Deft	Roh	Ave. Roh	Ratio to .0592[1]	Ave. Rohd	(6)[2] / (4)
Education of husband	None	.255	1.684	.0186	.1212	2.05	.0101	.111
	Primary	.548	1.727	.0201	.0615	1.04		
	Junior High	.081	1.453	.0112	.0410	0.69	.0053	.077
	Senior High +	.070	1.739	.0205	.0969	1.64		
Occupation of husband	Farmer	.219	2.437	.0509	.1474	2.49	.0208	.189
	Labr.&Opertv.	.202	2.002	.0310	.0726	1.23		
	Skilled	.149	1.951	.0289	.0733	1.24	.0041	.065
	White Collar +	.359	1.872	.0258	.0525	0.89		
Income of family (1000 NT)	0-23.9	.154	4.171	.1987	.1765	2.98	.0211	.160
	24.-35.9	.172	1.445	.0132	.0868	1.47		
	35.-47.9	.172	1.807	.0274	.0639	1.08	.0044	.067
	48. +	.303	2.476	.0621	.0671	1.13		
Ave. for 12 classes			2.064	.0424	.0884	1.494	.0110	.112
Children ever born	0-1	.147	1.221	.0050	.0671	1.13	.0036	.054
	2	.172	1.122	.0026	.0667	1.13		
	3	.239	0.987	-.0002	.0613	1.04	.0025	.036
	4 or more	.396	1.429	.0105	.0766	1.29		
Marriage duration	0-4	.228	1.139	.0031	.0622	1.05	.0031	.049
	5-9	.267	0.874	-.0024	.0647	1.09		
	10-19	.386	1.038	.0008	.0741	1.25	-.0001	-.001
	20 +	.058	1.037	.0008	.0936	1.58		
Age of wife	19-24	.189	1.150	.0032	.0554	0.94	.0014	.022
	25-29	.252	1.187	.0041	.0715	1.21		
	30-34	.260	1.169	.0037	.0678	1.14	.0006	.008
	35-42	.255	0.892	-.0021	.0733	1.24		
Ave. for 12 classes			1.104	.0024	.0694	1.174	.0019	.028
Ave. for 24 classes					.0790	1.334	.0064	.070

1. 0.0592 is the average roh for the 40 characteristics on the total sample (see bottom of Col. 4 of Table 1).
2. In calculating the ratio, the mean of the two entries in col. 4 is used.

average roh for the 24 subclasses is 0.0790, and the ratio 0.0790/0.0592 1.334 measures the average increase over the roh value based on the total sample.

3.4 Results for differences between subclass means

We have computed roh values for the difference of each of 2 pairs in each set of 4 subclasses, for each of the 40 characteristics. The averages over the 12 values are shown in col. 7 of Table 1, where rohd is the roh for the difference. These rohd values are substantially lower than the corresponding subclass values. The individual ratios (not shown) of values in column 6 to column 4 vary considerably around their average of .095. A better average is the ratio of means: .00652/.0790 = .083. The individual ratios range most from 0.30 to 0.00, except from some trivial cases near the bottom of the table, where negative values appear. We have also found in many other studies positive but smaller effects for differences than for the corresponding subclasses. The effects of covariance between subclasses seem unusually strong in this design. Consequently, the effects of clustering of differences though still present, are considerably reduced. In column 6 of Table 2 are shown roh values for differences of pairs of subclass means. Each of the 12 entries represents an average over the 40 variables of Table 1. Note the great reductions in design effects due to positive covariances in clusters. The ratios of the average rohs is .0064/.0790 = 0.081.

4. Highlights from other surveys

The 1971 and 1973 South Korea fertility studies provided an opportunity to study sampling errors for the same characteristics at two points in time. At first glance it seemed that the roh values in 1973 were considerably smaller than

those in 1971. The average roh value for some 40 characteristics was 0.049 in 1971 and 0.033 in 1973. However, when we examined only the subset of characteristics which were common to both surveys the average roh values were 0.037 in 1971 and 0.030 in 1973. In this subset the design effects are 3.85 and 2.02 respectively because the average cluster size in 1973 was much smaller than in 1971. This is an example of why we argue for portability in terms of roh rather than deft. The range of roh values in the South Korean fertility surveys was 0 to 0.2.

A fertility survey of Malaysia was conducted in 1969 and yielded 2,950 interviews with women involved in two large family planning programs. The sample was drawn after stratification into rural and urban areas. It was found that the design effects were far larger in the rural than in the urban areas. For 29 variables, the average deft's for the rural and urban areas were 1.92 and 0.99 respectively. The average roh for rural areas was 0,046. In the urban areas there was no clustering since the respondents were selected individually from lists of names. The range of roh values for the total sample was 0.02 to 0.05.

Arranging the characteristics by size of roh revealed two striking results. The characteristics "proportion using NFPB clinic," "proportion Malay" and "proportion with farmer husband" produced abnormally large sampling errors (deft's of 4.06, 2.65 and 2.58 and roh's of 0.36, 0.14 and 0.13 respectively). The first is explained by the fact that women in a given cluster either attended one type of clinic or the other. (This variable could have been an appropriate stratification variable.) The second result suggests that ethnicity is a highly clustered variable in Malaysia. The third result is due to the fact that clusters follow geographical boundaries with diverse densities of farmers.

Another result gleaned form the Malaysia survey is that subclasses that approximate crossclasses produce different sampling errors than do subclasses that are segregation classes. Over 5 pairs of crossclasses (e.g., income, age, marital status) the average roh across 14 characteristics was 0.0318, which has a ratio of 1.15 to the average roh for these characteristics on the total sample. On the other hand, if we consider the segregation classes (e.g., type of clinic, ethnicity, rural-urban birth and farmer-nonfarmer occupation) the average roh is 0.0750.

5. Summary of Results from Eight Surveys

For each survey sampling errors were computed for about 30 to 40 characteristics. This was done in each survey for means based on the entire sample and on about 24 subclasses and for differences between about 12 pairs of subclass means. The great range across different variables in values of roh in each of the surveys is the most important result. The roh values have an effective hundredfold range in each survey from about 0.001 to 0.002 to about 0.1 or 0.2.

Some differences between types of variables can be detected on each survey in Table 3. However these differences are not consistent and are also marked by considerable sampling variability. Socio-economic variables appear noticeably high for Korea and Peru. Demographic background variables tend to be near the lower end for all surveys. Attitudes and birth preferences appear high though more often in the lower half with roh values mostly from 0.005 to 0.05. The ranges within types (not shown) seem to be factors of about 5 to 10. They are considerably less than the range of 50 or 100 for rohs of all variables within surveys. Thus the typing of variables seems an effective and simple way to reduce our level of ignorance.

The individual computations of rohs for each characteristic/subclass combination are subject to great variability. But the average roh for each characteristic computed over several subclasses is quite stable. We refer to subclasses that are approximately crossclasses (more or less evenly distributed in the sample clusters). Other kinds of subclasses, those that are very unevenly distributed in sample clusters, need special considerations.

Table 3 summarizes a vast body of computations over the eight surveys. Since the variables included had not been coordinated initially, it is comforting that some very useful stabilities may nevertheless be drawn from them. The average values of overall rohs (first row) varies from .024 to .063. This stability is quite good, considering the diversity of variables and sample designs. It is helpful for choice of sample designs, since accepting .04 or .05 for roh would not badly mislead one. For fertility experience and demographic background variables, the roh values are lower and more stable, .011 to .038. For general attitudinal variables the roh values are very high for Taiwan and Peru and fertility preferences are also high in Taiwan. It would be interesting to investigate how much

TABLE 3

Summary of Average Roh's for Eight Surveys [a]

STATISTIC	SAMPLE SURVEY: South Korea 1971	South Korea 1973	Taiwan	Peru	Malaysia	United States 1960 Whites	United States 1970 Whites
A. ROH'S FOR TYPES OF VARIABLES FOR TOTAL SAMPLE (Number of characteristics below roh values)							
1. All Characteristics	.050	.033	.059	.063	.045	.024	.037
	40	39	40	29	29	9	36
2. Fertility Experience	.016	.009	.014	.034	.025	.011	.019
	11	6	9	8	3	4	6
3. Contraceptive Practice	.047	.021	.030	.054	.022	.043	.029
	9	11	9	8	3	2	8
4. Fertility Preferences	.023	.024	.072	---	.028	.025	.019
	6	8	11	0	3	2	6
5. Attitudes	.028	.026	.145	.094	.017	---	.061
	2	3	8	1	2	0	16
6. Socio-economic Variables	.128	.081	.016	.126	.045	---	---
	9	8	2	7	12	0	0
7. Age, Marriage (demographic background)	.014	.025	.025	.024	.010	.039	.105[b]
	3	3	1	5	2	1	1
B. ROH'S FOR SUBCLASSES AND FOR DIFFERENCES							
Number of Characteristics	40	39	40	20	14	9	36
Number of Subclasses	23	22	24	10	20	8	24
8. Roh's for Total Sample	.050	.033	.059	.056	.028	.024	.037
9. Roh's for Subclasses	.059	.044	.079	.065	.032[c]	.048	.052
10. Ratio of Subclass/Total (9)/(8)	1.19	1.36	1.33	1.15	1.15[c]	2.00[d]	1.41
11. Differences of Means	.0060	.0000	.0065	.0170	.0300[c]	.0130	.0050
12. Ratio of Difference/Subclass (11)/(9)	.100	.000	.083	.026	.210[c]	.270	.096
C. COMPARISONS OF SUBCLASSES: SOCIO-ECONOMIC (SE) VERSUS CROSSCLASSES							
13. SE as Characteristics	.076	.092	.042	.105	---	---	.122
14. Others as Characteristics	.006	.007	.002	.015	.037	---	.020
15. SE Subclass Base	.063	.040	.088	.073	---	---	.063
16. Others as Subclass Base	.057	.038	.069	.063	.032	---	.047

[a] The eighth survey pertaining to blacks in 1970 was unreliable due to sample design and small sample size.

[b] Results unreliably high for unknown reasons.

[c] Results are for 10 crossclasses only.

[d] This result is based on 8 subclasses and removing one of them reduces the ratio to 1.15.

of these high roh values are due to homogeneity of the respondents in compact clusters, or how much of the effects of interviewer variance of response from large workloads. The high roh values for socio-economic variables in Peru and South Korea have implications for sample designs, as well as for sociological studies of their sources.

When we separate socio-economic subclasses from others we regularly note considerable differences between the two groups, when these are computed as characteristics based on the entire sample (rows 13 and 14). However, when used as subclasses (rows 15 and 16) the differences between the two sets of subclass roh's (averaged over all characteristics) are not great, say 1.2 versus 1.4. It is the characteristics, much more than the subclass, that are the sources of variability in sampling errors.

The ratio of the rohd's for difference to the average roh's for subclass means (rows 11 and 12) is not stable. In all cases the reductions due to covariances between clusters are substantial. The central value may be 0.1 and 0.2.

6. Strategies for Large-Scale Calculation, Summarization and Presentation of Sampling Errors
   (1) Paired selection considerably simplifies sampling error calculations.
   (2) The coefficient of variation of cluster size should always be calculated and inspected before the results of sampling error calculations are published, since the approximate formula for var(r) requires $cv(x) < 0.2$.
   (3) Codes identifying the primary sampling units and the strata must be included together with the data. Our experience has been that these codes are seldom readily available.
   (4) Sampling errors should be calculated for the entire sample for many variables. We think it inadequate to single out a few critical survey variables or several categories of one variable. Rather than exhausting all categories for a few variables, more variables should be used, each one for one or a few categories. Variability between variables is generally greater than between categories within variables. This is especially true for characteristics, but also for subclass variables. The range of variables should parallel the aims of the survey, of its analysts and of its users. Also, it should aim to cover the range of design effects.
   (5) The variables should be separated into a few groups within which the sampling errors are expected to be relatively similar.
   (6) Sampling errors should be computed for many characteristics each based on a moderate number of subclasses. Sampling errors, particularly roh's, were found subject to greater diversity across characteristics than across subclasses. Subclass results should be compared to the results obtained for the total sample.
   (7) Most of the needed subclasses tend to approximate crossclasses. However, partially segregated subclasses, if important, should also be investigated.
   (8) In choosing subclass categories a range of subclass sizes should be selected to obtain empirical evidence of the effect of subclass size on deft and roh.
   (9) All chosen characteristics should be analyzed by all chosen subclasses (rather than using different subclasses for each characteristics). This yields a symmetrical table and averaging can be done over both subclasses and characteristics. However, other designs may be used, especially for a larger number of subclasses.
   (10) Sampling errors should be computed for the difference of means of pairs of subclasses. For many subclass variables one or two pairs usually suffice. These results should be compared with the individual results for each of the two subclasses.
   (11) Sampling error results should be preserved and publicized for the use of survey designers who would find such data useful in the design of future surveys.

In addition to the 40 characteristics that we treated as "dependent," we also computed roh values for 24 variables later used for subclass analysis. Here a clear dichotomy emerged. The 12 characteristics based on demographic variables had roh values under 0.005 (Table 2, col. 3). However, the 12 socioeconomic characteristics had roh values 0.01 to 0.20. Within the two classes of characteristics there is variation, but much of it is too haphazard to be of general use.

MORE DALLYING WITH CPS DESIGN EFFECTS

Thomas N. Herzog, Social Security Administration*

This paper describes an analysis of various sets of design effects constructed from the Census Bureau's March 1973 Current Population Survey (CPS). The paper is divided into five parts. In the first part we present the basic definitions, a discussion of our earlier results, and some limitations on the calculations to be performed. The second part is an investigation of some conjectures (of Kish and Frankel [1]), as they pertain to the CPS. In order to produce summary descriptors of collections of design effects, we consider, in part three, various schemes of averaging design effects. Since these schemes all appear to be unsatisfactory, in part four we propose an alternative type of summary descriptor based on the concept of empirical Stein estimation (see, for example, [2]). Finally, part five consists of a few brief concluding remarks.

1. INTRODUCTION AND BACKGROUND

1.1 Design effects.--Standard statistical methods have been developed under the assumption of simple random sampling (SRS). Although the independence of sample elements is often assumed, it is seldom realized in large complex surveys. As a result, practitioners [3, 4, 5] suggest alternative methods, such as jackknifing or the use of balanced repeated replication, for calculating sampling errors in complex surveys. Design effects [6] are essentially just measures for comparing such estimates of the "actual variance" to those computed under the SRS hypothesis.

In particular, for a given statistic X, we define the design effect of X, $\delta(X)$, by

$$\delta(X) = \frac{VAR(X)}{\sigma^2(v)} \tag{1.1}$$

where VAR(X) is the (expected) variance of X for the actual complex survey, and $\sigma^2(X)$ is the expected variance of X which would have been obtained by selecting, with replacement, a simple random sample of exactly the same size from the entire population surveyed. For example, if P is the actual proportion of items in a population with a given characteristic and n is the sample size, then the SRS variance of the usual estimator, $\hat{P}$, of P is

$$\sigma^2(\hat{P}) = P(1-P)/n. \tag{1.2}$$

The design effect, δ, is a measure of the impact on the actual variance of the complexity of the sample design relative to that of simple random sampling; in other words, δ summarizes the composite effect on the variance of such things as the number and nature of the selection at each stage of the sampling process, the extent of pre- and post-stratification, and the ultimate cluster size. We will use δ to refer to population values and $\tilde{\delta}$ to sample values.

1.2 Summary of previous results.--In a paper delivered at the 1976 Annual Meeting of the American Statistical Association [7], Fritz Scheuren and I presented an empirical study that considered:

(i) various methods of calculating individual design effects for proportions, and

(ii) various methods of averaging these individual design effects.

The principal conclusions of that work were:

(i) Each of the (asymptotically-equivalent) design effect estimators considered produced essentially the same value. (This suggests that, for our data, each estimator considered was equally good.)

(ii) Different methods of averaging these design effects produced substantially diverse summary statistics.

The results on averaging methods in [7] warranted further examination and led directly to the present effort.

1.3 Statistics considered.--In last year's paper we considered design effects for CPS STATS units by race of the unit head.1/ Within each racial group, design effects were calculated separately for five different classifiers: type of unit, total unit size, total earnings of unit, total social security benefits of unit, and total income of unit. The asymptotically unbiased estimators whose design effects we examined were

(i) $\hat{P}(W)$, the proportion of whites in a given category, and

(ii) $\hat{P}(B)$, the proportion of nonwhites in a given category. (Hereafter, we will refer to nonwhites as "blacks.")

In the present paper, we re-examine these design effects, as well as those of

(iii) $\hat{D} = \hat{P}(W) - \hat{P}(B)$, the difference in the proportion of whites and "blacks" in a given category,

(iv) Yule's Q, and

(v) the cross-product ratio, denoted by C.

The last two statistics measure the association between the variables race (white or black) and inclusion (or exclusion) in a given category. In particular [9, p. 539], if we have the table of observed frequency counts

	White	Black
In category	a	b
Not in category	c	d

then Yule's Q and the cross-product ratio C may be estimated as

$$(1.3) \qquad \tilde{Q} = \frac{ad-bc}{ad+bc}$$

and

$$(1.4) \qquad \tilde{C} = \frac{bc}{ad} = \frac{1-\tilde{Q}}{1+\tilde{Q}}$$

This definition of the cross-product ratio is the reciprocal of the usual one. We use the symbol $\delta(W)$ to denote the set of design effects for the proportion of whites. Similarly, we use $\delta(B)$, $\delta(D)$, $\delta(Q)$, and $\delta(C)$ to denote, respectively, the set of design effects for the proportion of blacks, the difference in proportions, Yule's Q, and the cross-product ratio.

1.4 Replicate estimators of design effects.--There are, of course, many ways to construct estimators of design effects. In parts 2 and 3, we confine our attention to jackknife estimators which pertain when the sample may be separated into a number, say r, of independent, identically designed subsamples or replicates. The "replicates" employed in our study are the eight rotation panels of the March 1973 CPS.2/ For a particular set of design effects, say $\delta(W)$, we will basically employ estimators of the form

$$(1.5) \qquad \hat{\delta}(W) = \frac{\hat{VAR}(W)}{\tilde{\sigma}^2(W)}$$

where

$\hat{VAR}(W)$ is the jackknife estimator of the actual variance of $\hat{P}(W)$ and $\sigma^2(W)$ is an asymptotically unbiased estimator of the SRS variance of $\hat{P}(W)$.

Formulas for computing the SRS variance estimates considered here appear in [9] under the assumption that the sampling of blacks and whites is carried out independently. In particular, our estimate of $\sigma^2(W)$ is

$$(1.6) \qquad \tilde{\sigma}^2(W) = \tilde{P}(W)\left[1-\tilde{P}(W)\right]/n(W),$$

where n(W) denotes the total number of whites surveyed.

Estimates of all of the actual variances and some of the design effects are obtained by using the jackknifing technique. As in last year's paper, jackknifing is also used to calculate the standard errors of all the design effects considered.

1.5 Some limitations.--Because the same sample of PSU's is common to all rotation panels, it is not possible to use the panels to estimate the between-PSU component of the CPS variance. Consequently, the "design effects" considered here relate only to the within-PSU component of the estimators. It might be mentioned, parenthetically, that for each statistic discussed in this paper, the within-PSU component probably accounts for at least 90 percent of the total variation.

The Census Bureau constructs all eight rotation panels in the same way. As already stated, we are using these 8 panels as the r=8 replicates. Consequently, there is considerable variation (from 1 to 8) between panels (i.e., "replicates") in the number of times each of the interviewees is surveyed prior to and including the March 1973 interview.

Differences in the method of conducting the interviews also exist from panel-to-panel. Initially, the questions are asked in person; but, in the later panels, most of the surveying is done by telephone. The net effect of these and other factors [10] is to alter the response patterns from panel-to-panel so that the panels cannot be assumed to be a priori identically distributed. The influence of these panel differences on the statistics under consideration here is not known.3/ When we began this work, we implicitly assumed that such panel effects, if any, would be small enough to ignore. This was in part, a reflection of our, perhaps misplaced, confidence in the nature of the raking ratio estimation procedures employed.4/ Project plans call for a repetition of the present calculations using a random group estimator (described in [12]) that would not be subject to "panel biases."

2. AN EMPIRICAL COMPARATIVE INVESTIGATION OF SOME DESIGN EFFECT ESTIMATORS

2.1 Kish-Frankel conjectures.--This part of the paper is inspired by some conjectures of Kish and Frankel [1; p. 13]. Having defined $\bar{Y}$ as the mean of the vector of statistics $\underline{Y}$ and A as a complex function of $\underline{Y}$, we may list the Kish-Frankel conjectures as

(i) $\delta(A)>1$. In general, the population values of the design effects of complex statistics tend to be greater than 1.

(ii) $\delta(A)<\delta(\bar{Y})$. The design effect of the mean $\bar{Y}$ of a statistic $\underline{Y}$ tends to be greater than those of complex functions of $\bar{Y}$.

(iii) $\delta(A)$ is related to $\delta(\bar{Y})$. For variates with high $\delta(\bar{Y})$, values of $\delta(A)$ tend also to be high.

(iv) $\delta(A)$ tends to resemble the design effect for differences of means.

(v) $\delta(A)$ tends to have observable regularities for different statistics.

A simple model of the above would be

(2.1) $\delta(A_g) = 1 + f_g[\delta(\bar{Y}) - 1]$, with $\delta(\bar{Y}) > 1$

and $0 < f_g < 1$ and f_g

specific to the variables and statistic denoted by g.

The calculations in this part of the paper are performed for both the original five "basic" sets of design effects and for "high-proportion" sets which are created by deleting from the basic sets those categories in which either the proportion of whites is less than 2% or the proportion of blacks is less than 5%.

2.2 Conjecture (i).--The first conjecture we examine is that the values of the design effects tend to be larger than 1. For the basic sets, each composed of 63 individual design effects, we find that 74.60% of the elements of $\tilde{\delta}(W)$ are greater than 1. However, none of the other sets of design effects, including $\tilde{\delta}(B)$, shares this property; only about 50% of these values tend to be larger than 1.

For the high-proportion sets, each composed of 32 individual design effects, 75% of the elements of $\tilde{\delta}(W)$ are larger than 1. Moreover, for the other four high-proportion sets, the percentage of values greater than 1 increases, although remaining somewhat below that of $\tilde{\delta}(W)$.

2.3 Conjecture (ii).--We next compare the values of the individual design effects of each set to the corresponding values of each of the other sets of design effects. For the basic sets, we find that the elements of $\tilde{\delta}(W)$ tend to be larger (in about two-thirds of the cases) than the corresponding elements of the other sets of design effects. For example, 63.49% of the elements of $\tilde{\delta}(W)$ exceed the corresponding values of $\tilde{\delta}(B)$. For the other four basic sets, no one set particularly dominates any other. For the high-proportion cases, the values of $\tilde{\delta}(W)$, again, tend to be the largest. The values of $\tilde{\delta}(B)$ tend to be less than those of the three complex statistics; the values of $\tilde{\delta}(D)$ and $\tilde{\delta}(Q)$ are both generally less than those of $\tilde{\delta}(C)$.

In light of conjecture (ii), it is not surprising that the values of $\tilde{\delta}(W)$ dominate the values of the other sets; however, it is, at least at first glance, surprising that the values of $\tilde{\delta}(B)$ tend to be smaller than the values of the sets corresponding to the three complex statistics.

2.4 Conjectures (iii), (iv), and (v).-- We next examine the correlation coefficient of each pair of sets of design effects. We find, for the basic sets, $\tilde{\delta}(W)$ is positively correlated with $\tilde{\delta}(D)$ and negatively correlated with $\tilde{\delta}(B)$, $\tilde{\delta}(Q)$ and $\tilde{\delta}(C)$, the value of each of these four correlation coefficients being relatively close to zero; i.e., .0238, 0.0871, -.0502, and -.0495, respectively. For the high-proportion sets, $\tilde{\delta}(W)$ is, again, negatively correlated with each of the other four sets, but here the magnitude of each of these correlation coefficients is relatively large.

Excluding $\tilde{\delta}(W)$, the remaining four sets are very strongly positively correlated. This is not surprising. Since about eight times as many whites are surveyed as blacks, the blacks account for roughly 85% of the variance of the difference in proportions. Furthermore, since

$$(2.2) \qquad 0 < \frac{2P(W)\cdot P(B)}{P(W) + P(B)} < 1,$$

we may write Q as

$$(2.3) \qquad Q = \frac{P(W) - P(B)}{P(W) + P(B)}\left[1 + \frac{2P(W)P(B)}{P(W) + P(B)} + \left(\frac{2P(W)P(B)}{P(W) + P(B)}\right)^2 + \ldots\right].$$

Also, since $-1 \le Q \le 1$,

we may write C as

$$(2.4) \qquad C = (1 - 2Q + 2Q^2 - 2Q^3 + \ldots).$$

So C may be approximated by 1-2Q, especially when the magnitude of Q is small. Thus, C is approximately a linear function of Q, and Q is approximately a linear function of the difference in proportions. It is, therefore, reasonable that the design effects for $\tilde{D}$, $\tilde{Q}$ and $\tilde{C}$ tend to be nearly equal and are so high correlated.

Thus, considering the difference in proportions, Yule's Q and the cross-product ratio as complex functions of the proportion of blacks, we have an even stronger result than conjecture (iii); namely, that the design effects of (certain) complex statistics are highly-correlated with the design effects of the proportion of blacks.

3. ORIGINAL DESIGN EFFECT AVERAGING SCHEMES

In this part of the paper we reconsider the averaging schemes employed in our earlier paper. These schemes are applied to a number of sets of design effects not considered previously. Our goal here is to discover a good summary descriptor of sets of design effects.

In our earlier paper, we considered four types of "averages"--the median and three means (arithmetic, harmonic, and geometric). We also employed three distinct weighting schemes--uniform weighting, weighting by the reciprocal of the estimated simple random sampling (SRS) variances, and weighting by the reciprocal of the estimated SRS relvariances. Applying these 4 x 3 = 12 averaging schemes to our five basic sets of design effects, we obtain the data of table 1.

The results of two additional averaging schemes are also shown in table 1. The first scheme, suggested by Kish, is the square of the average

of the square roots of the individual design effect estimates. Kish [6, p. 578] prefers this scheme. The second, referred to as the overall ratio average, is the average of all of the individual estimated actual variances divided by the corresponding average of the estimated SRS variances.

Last summer we were rather surprised that our 14 averaging techniques produced such diverse numerical results as those displayed in table 1. Consequently, we have since examined these calculations in much greater detail.

The most striking phenomenon concerned the two sets of non-uniform weighting schemes. In several instances, a relatively small number of the individual classes under consideration accounted for the predominant share of the weight. Consequently, in these cases, the values of the vast majority of the design effects of a particular set had almost no influence on the value of the summary statistic produced.

It is instructive at this point to consider a specific case : the relvariance weighting scheme applied to the estimators of the design effects of the proportion of blacks. In this particular instance, three of the 63 classes account for over 70% of the weight. These three classes are:

(i) STATS units receiving no social security benefits (25.24%);

(ii) STATS units having total earnings of less than $10,000 (23.24%); and

(iii) STATS units having total income of less than $10,000 (22.11%).

In our earlier paper we also attempted to partition the sets of design effects into subsets which would be more homogeneous. In particular, the estimated design effects for the proportion of whites and blacks were partitioned into three or four groupings according to the estimated value of the corresponding proportion. This procedure narrowed the range of the averages substantially; however, the numerical differences among the various averaging schemes were still "uncomfortably" large.

4. STEIN ESTIMATORS OF DESIGN EFFECTS

In light of the diverse results of the averaging schemes just presented, we decided to consider another method of constructing a summary descriptor of a set of design effects. The approach taken is discussed in Geisser [13, 14] and is based upon an empirical Stein-type estimator. Geisser's approach is of a heuristic, ad hoc nature. Its justification lies in whether or not it works in a given situation. We believe that such a scheme can be profitably applied to our data.

The Stein estimator, as originally formulated [2], requires a number of stringent assumptions, some of which are clearly not valid in the present situation. On the other hand, Efron and Morris [15], among others, argue that the violation of these assumptions does not necessarily diminish the estimator's usefulness.

In the remainder of the paper we will discuss these issues as they pertain to our CPS data. The reader should keep in mind that we are only attempting to do some "dallying" with a few sets of design effects and are not attempting to resolve any of the outstanding theoretical issues concerning the general applicability of empirical Stein estimation.

4.1 Original Stein estimator.--We present here a brief description of Stein estimation. We, first, let $j=1,\ldots,J$ and $k=1,\ldots,K$ where $K \geq 3$. For a given collection of parameters $\{\Theta_j\}$, we assume that the random variables $\{X_{kj}\}$ are independent and normally distributed with means $\{\Theta_j\}$ and common variance σ^2. In this setting, we define the Stein estimator of Θ_j to be

$$(4.1) \qquad \hat{X}_j = \mu X_{..} + (1-\mu)X_{.j}$$

where $X_{.j} = \frac{1}{K} \sum_{k=1}^{K} X_{kj}$ and

$$(4.2) \qquad X_{..} = \frac{1}{J} \sum_{j=1}^{J} X_{.j}$$

The unknown parameter "μ" is such that

$$(4.3) \qquad C \leq \mu \leq 1.$$

James and Stein [2] have shown that for an appropriate choice of μ, the use of the estimator $\hat{X}_j$ produces, on the average, a smaller mean square error than the maximum likelihood estimator.

Following Geisser [13], we let $\min(\mu_1, 1)$ be an estimator of μ where

$$(4.4) \qquad \mu_1 = \frac{(JK-1)\, m_1}{(J-1)m_1 + (K-1)Jm_2}$$

$$= \left[\frac{(J-1)}{(JK-1)} + \frac{(K-1)J}{(JK-1)} \frac{m_2}{m_1} \right]^{-1}$$

with

$$(4.5) \qquad m_1 = \frac{1}{J(K-1)} \sum_{k=1}^{K} \sum_{j=1}^{J} (X_{kj}-X_{.j})^2 \text{ and}$$

$$(4.6) \qquad m_2 = \frac{K}{J-1} \sum_{j=1}^{J} (X_{.j}-X_{..})^2 .$$

4.2 Stein estimation of design effects.--It now remains to relate the above formulation to the problem at hand. To limit the amount of computation involved, we restrict our attention to

$\tilde{\delta}(W) = \{\tilde{\delta}_j(W)\}$, the set of computed design effects for the proportion of whites. In addition, we only consider the harmonic and Kish (unweighted) averaging schemes, as these produced quite diverse results when applied to the individual $\tilde{\delta}_j(W)$. (See table 1.)

Our first approach is to replace

(i) the $\{\theta_j\}$ by $\{\delta_j(W)\}$, the actual (expected) values of the white design effects,

(ii) the $\{X_{.j}\}$ and $\{\hat{X}_j\}$ by $\{\tilde{\delta}_j(W)\}$ and $\{\hat{\delta}_j(W)\}$ respectively, and

(iii) $X_{..}$ by $\tilde{\delta}_.(W)$, an appropriate (i.e., harmonic or Kish) average of the set of individual design effects.

Noting that m_1 equals K times the average of the usual estimators of the variances of the $\{X_{.j}\}$, we can write the "m_1" and "m_2" values corresponding to the $\tilde{\delta}_j(W)$ as

$$(4.7) \qquad \tilde{m}_1 = \frac{K}{J} \sum_{j=1}^{J} \sigma^2\left(\hat{\delta}_j(W)\right)$$

and

$$(4.8) \qquad \tilde{m}_2 = \frac{K}{J-1} \sum_{j=1}^{J} \left(\tilde{\delta}_j(W) - \tilde{\delta}_.(W)\right)^2 .$$

Hence, we have

$$(4.9) \qquad \frac{\tilde{m}_2}{\tilde{m}_1} = (1-1/J)^{-1} \left[\frac{\text{Mean squared deviation of the } \tilde{\delta}_j(W) \text{ from the overall average}}{\text{Mean of the estimated variances of the } \{\hat{\delta}_j(W)\}}\right].$$

This ratio can be evaluated by using the desired overall average $\tilde{\delta}_.(W)$, together with the $\{\tilde{\delta}_j(W)\}$, and the corresponding jackknife estimated variances.

It is now possible to estimate μ_1 by substituting $\frac{\tilde{m}_2}{\tilde{m}_1}$ for $\frac{m_2}{m_1}$ in equation (4.4) and choosing suitable values for J and K. The choice of K=8 and J=63 is not optimal because it ignores the facts that the $\{\tilde{\delta}_j(W)\}$, while based on 8 independent replicates, are not sample means, and that the design effect for each of our 63 categories is not independent of those of the other categories.

Several ad hoc "solutions" to this selection problem were considered.5/ The one which seems most reasonable to us is to note [13] that

$$(4.10) \qquad \mu = \min(\mu_1, 1) \geq \min\left(\frac{m_1}{m_2}, 1\right)$$

and to simply choose $\min\left(\frac{\tilde{m}_1}{\tilde{m}_2}, 1\right)$ as our estimate of μ. Computing a value for μ in this manner, we find that it equals 1 for both the Kish and harmonic averaging schemes. This result was rather disappointing in that it leaves us exactly where we were at the end of section 3, with different averaging schemes producing diverse numerical results and no way to choose among them.

We suspect that the value of $\mu = 1$ results from the heteroscedastic nature of the variances of the $\hat{\delta}_j(W)$. In order to "eliminate" this source of concern, we redefine $\tilde{m}_1$ and $\tilde{m}_2$ as

$$(4.11) \qquad \tilde{\tilde{m}}_1 = \frac{K}{J} \sum_{j=1}^{J} \frac{\sigma^2\left(\hat{\delta}_j(W)\right)}{\tilde{\delta}_j^2(W)}$$

and

$$(4.12) \qquad \tilde{\tilde{m}}_2 = \frac{K}{J-1} \sum_{j=1}^{J} \frac{\left(\tilde{\delta}_j(W) - \tilde{\delta}_.(W)\right)^2}{\tilde{\delta}_.^2(W)} .$$

Estimating μ this time as $\min\left[\frac{\tilde{\tilde{m}}_1}{\tilde{\tilde{m}}_2}, 1\right]$, we produce μ values of 0.5040 and 0.8885 for the harmonic and Kish schemes, respectively. Using these values of μ, we may re-estimate the white design effects by

$$(4.13) \qquad \tilde{\tilde{\delta}}_j(W) = \mu\tilde{\delta}_.(W) + (1-\mu)\tilde{\delta}_j(W).$$

Our next task is to compare the two sets of design effects calculated from equation (4.13). Our approach involves calculating, for each averaging scheme, the nominal length of the symmetric 95% confidence interval of the proportion of whites in each of the 63 categories. This is done under the assumption 6/ that

$$E\,\hat{\hat{\delta}}_j(W) = \delta_j(W),$$

where $\hat{\hat{\delta}}_j(W)$ is the estimator corresponding to $\tilde{\tilde{\delta}}_j(W)$.

It can be shown that under regularity conditions the length of the j-th interval is proportional to

$$(4.14) \qquad \sqrt{\tilde{\tilde{\delta}}_j(W)}\cdot t(.95, DF(J)),$$

where $t(.95, \tilde{\tilde{DF}}(j))$ is the length of a symmetric 95% confidence interval for a random variable having a Student's t distribution with $\tilde{\tilde{DF}}(j)$ degrees of freedom. $\tilde{\tilde{DF}}(j)$ is determined from

$$(4.15) \qquad \tilde{\tilde{DF}}(j) = 2/\text{relvariance}\ (\hat{\hat{\delta}}_j(W)).$$

Using this procedure, we find that, under Stein estimation, the length of the average confidence interval corresponding to the harmonic mean is 100.8% of that of the Kish mean. This compares to a value of 90.52% for the corresponding (unadjusted) averaging schemes of section 3. (This last quantity is simply the square root of the ratio of the harmonic average of the white design effects to that of the Kish average.)

Since the Stein estimation procedure has produced confidence intervals whose average lengths are more nearly equal under the harmonic and Kish schemes, the Stein technique appears to compare favorably, at least for our data, to the unadjusted overall averaging schemes of section 3. It should be pointed out, however, that we have gone only a very small part of the way towards applying Stein estimation to design effects. The chief difficulty, not addressed in this paper, is that the confidence intervals are, in general, biased; hence, in some situations, they could be very badly mis-estimated.

5. SOME CONCLUDING REMARKS

First and foremost, we must, again, emphasize that we have performed an empirical examination of the data of a single sample survey. In addition, we have only considered a very limited number of statistics. In general, the analysis described in part two confirms the conjectures of Kish and Frankel [1; p. 13] .

As in our earlier paper, the averaging schemes discussed in part 3, unfortunately, produced widely diverse results. This may be because the sets of design effects considered were not sufficiently homogeneous for some or all of the averaging methods. On the other hand, the empirical Stein estimation scheme described in part four produced somewhat better results; i.e., the lengths of the confidence intervals were on the average more nearly equal. These improved results were, however, obtained by the application of an ad hoc technique to a single set of data. Thus, our evidence in support of the Stein estimation scheme is not exactly overwhelming.

There is no doubt that much theoretical work is needed to "resolve" the issues raised here. As Kish and Frankel [1, p. 13] suggest, such theory will need to be buttressed by empirical results. We, therefore, encourage others to do some "dallying" with their own favorite sets of design effects, as we will continue to do ourselves.

ACKNOWLEDGEMENTS AND FOOTNOTES

*The author would like to thank Fritz Scheuren for his advice, encouragement, and moral support in the preparation of this paper. The author is also indebted to Professor B. V. Sukhatme for several helpful comments and to Roland Minton for computer-programming assistance. Typing assistance was provided by Joan Reynolds and Helen Kearney.

1/ A "STATS" unit is a group of individuals in a CPS household who would generally be considered to be interdependent under social insurance programs. The STATS unit concept is defined in [8] .

2/ See subsection 1.5 below for the limitation imposed by this use of rotation panels as replicates.

3/ It should be pointed out, however, that to the extent that there are any panel differences, these would lead to an increase in the expected value of the estimated design effects.

4/ The estimator being used is described in [11] where it is referred to as the "intermediate undercount raking weight."

5/ One such "solution" involves letting

$$\mu = Km_1/\left[(K-1)m_2 + m_1\right] \quad \text{with } K=7.$$

6/ We have some results for the more realistic and interesting case when $E\hat{\delta}_j \neq \delta_j(W)$, but these were not presented at the session and, in any case, are still incomplete.

REFERENCES

[1] Kish, L. and Frankel, M.R., "Inference from Complex Samples," Journal of the Royal Statistical Society, Series A, Vol. 137, No. 1, 1974, pp. 1-22.

[2] James, W. and Stein, C. H., "Estimation with Quadratic Loss Function," Proceedings of the Fourth Berkeley Symposium, Vol. 1, 1961, pp. 361-379.

[3] Miller, R. G., "The Jackknife - a Review," Biometrika, Vol. 61, 1974, pp. 1-15.

[4] Kish, L. and Frankel, M. R., "Balanced Repeated Replications for Standard Error," Journal of American Statistical Association, Vol. 65, 1970, pp. 1071-1094.

[5] McCarthy, P. J., Vital and Health Statistics, PHS Publication No. 1000, Series 2, Nos. 14 and 31 (especially No. 31, pp. 12-15).

[6] Kish, L., Survey Sampling, New York, Wiley, 1965, pp. 574-582.

[7] Herzog, T. N. and Scheuren, F., "Dallying with Some CPS Design Effects for Proportions," 1976 American Statistical Association Proceedings, Social Statistics Section, 1977, pp. 396-401.

[8] Projector, D.S., et al, "Projection of March Current Population Survey: Population, Earnings, and Property Income, March 1972 to March 1976," Studies in Income Distribution, DHEW Publication (SSA) 75-11776, 1974.

[9] Kendall, M. and Stuart, A., The Advanced Theory of Statistics, London, Griffin, Vol. 2, 1967.

[10] Bailar, B., "The Effects of Rotation Group Bias on Estimates from Panel Surveys," Journal of American Statistical Association, Vol. 70, 1975, pp. 23-30.

[11] Scheuren, F., "Methods of Estimation for the 1973 Exact Match Study," (unpublished working paper to appear in the series Studies from Interagency Data Linkages).

[12] Bounpane, P., "General Outline of the Steps in the Keyfitz Estimate and Variance Programs," U.S. Census Bureau, unpublished memorandum, November, 1974.

[13] Geisser, S., "The Predictive Sample Reuse Method with Applications," Journal of American Statistical Association, Vol. 70, No. 350, 1975, pp. 320-328.

[14] Geisser, S., "A Predictive Approach to the Random Effect Model," Biometrika, Vol. 61, No. 1, 1974, pp. 101-107.

[15] Efron, B. and Morris, C., "Data Analysis Using Stein's Estimator and Its Generalizations," Journal of American Statistical Association, Vol. 70, No. 350, 1975, pp. 311-319.

Table 1.--Selected methods of averaging CPS within-PSU design effects: Usual and jackknifed estimators of the design effects, standard error and coefficient of variation of averages

Item	Jackknife Estimator					Coefficient of Variation				
	Proportion of whites	Proportion of blacks	Difference in proportions	Yule's Q	Cross-Product Ratio	Proportion of whites	Proportion of blacks	Difference in proportions	Yule's Q	Cross-Product Ratio
Uniform unit weighting:										
Arithmetic	1.6200	1.2574	1.2384	1.2509	1.2160	.1345	.0714	.1124	.0930	.0977
Geometric	1.4243	.9964	1.0106	1.0183	.9945	.1513	.0653	.1244	.0908	.0932
Harmonic	1.2440	.7238	.3040	.8218	.7978	.1787	.1974	.1849	.1132	.1163
Median	1.4249	1.0292	1.0743	.9646	.9920	.1008	.1736	.1726	.1300	.1196
Weighting by reciprocal of estimated variances under simple random sample:										
Arithmetic	1.3819	1.2263	1.1231	1.2673	1.1046	.0631	.1208	.1515	.0950	.0852
Geometric	1.2443	.9156	.8868	1.0657	.9519	.1153	.0697	.1744	.1097	.1322
Harmonic	1.1276	.5615	.6765	.8997	.8071	.1891	.3291	.2565	.1254	.1433
Median	1.2664	.9474	.9326	1.0219	.9920	.1967	.2214	.2811	.1614	.3939
Weighting by reciprocal of estimated relvariances under simple random sample:										
Arithmetic	2.2627	.9146	1.3046	1.2259	1.2728	.3721	.1276	.2119	.1469	.1079
Geometric	2.0254	.8402	1.1722	1.0827	1.0668	.3197	.1789	.2620	.2102	.1077
Harmonic	1.7791	.7952	.9984	.9506	.8962	.2869	.2116	.2835	.2203	.1791
Median	2.2711	.7550	1.2466	1.0219	.9920	.5310	.3477	.4484	.3688	.1571
Kish approach	1.5201	1.1260	1.1230	1.1303	1.1033	.1504	.0625	.1122	.0889	.0922
Overall ratio average	1.9517	1.3103	1.3160	1.1570	.8377	.3375	.1017	.1133	.1128	.1214

Note: The values shown for proportions in this table differ somewhat from the corresponding values shown in last year's paper because, even though the same data set was used, the way we defined the categories was altered slightly.

DISCUSSION

B. V. Sukhatme, Iowa State University

One of the papers I am going to discuss is concerned with estimation of variance while the other two are concerned with presentation and analysis of sampling errors. These are important topics and have not received as much attention as they deserve. I wish to congratulate the authors of these papers for their contributions. I am also thankful to the chairman for giving me an opportunity to participate in the discussion.

I shall first consider the paper by Lawrence Cahoon. The paper emphasizes rightly the importance of estimating variances for State estimates from the Current Population Survey and discusses several procedures. Since, I am not fully conversant with the design of the survey, my comments may sometimes be in the nature of questions and may be even naive.

i) In regard to the design of the survey, it has been remarked that in the nonself-representing strata, in addition to drawing one PSU from each stratum another PSU was drawn from each pair of grouped strata independent of the first selection. Since the strata are not large, it is likely that the same PSU may get selected in the additional sample resulting in reduced precision. It may be desirable to investigate whether the reduction in precision is appreciable.

ii) It has been mentioned that a control was exercised in the selection of the PSU's to ensure that one PSU was chosen in every state and the district of Columbia. It is not clear whether proper allowance has been made in the estimation procedure.

iii) The author has considered the collapsed strata estimator of variance discussed by Hansen, Hurwitz and Madow (1953). Special cases of this estimator of variance have been considered by Cochran (1953) and Seth (1966). This is a biased estimator of variance. The bias consists of two terms. Assuming that the stratum variance $\sigma^2_{x_{gh}}$ is proportional to its size Agh, one of the bias terms reduces to differences in strata sizes Agh. The other bias term due to differences in strata characteristics is simplified by assuming that the variance for a group of strata is proportional to $X_g = \sum_h X_{gh}$. The two assumptions are clearly not the same. If simplification is the main consideration, several other assumptions are possible. Infact the bias term due to differences in strata sizes vanishes altogether if Agh are not used and we use the estimator suggested by Cochran or Seth. Using the available data, it may be worthwhile to investigate whether any one of the two assumptions made is at all satisfactory.

iv) The problem of estimating the variance with one unit per stratum has also been considered by Hartley, Rao and Kiefer (1969). In certain situations, their method may lead to smaller bias in variance estimation than the method of collapsed strata. It may be desirable to include this method in the investigation.

v) Three different methods of groupings have been considered. These are based on 1970 projected unemployment rate. It may be worthwhile to investigate grouping based on the total number unemployed. This may turn out to be a different grouping and perhaps more efficient. However, the main concern is the use of the 1960 data both for grouping purposes and evaluation of the bias. The results obtained in respect of bias cannot as such be taken at their face value.

vi) On intuitive grounds, it appears that method III should result in minimum bias and this is confirmed by the numerical results given in Tables 2, 4 and 6. This is a positive result and needs further confirmation through numerical investigations of the type carried out in this paper.

vii) To evaluate the three grouping methods in respect of their mean square error, an approximation to the variance of the variance estimator has been obtained by using the formula developed by Hansen, Hurwitz and Madow (1953). This formula is derived under several restrictive assumptions which appear at variance with those made in evaluating the bias. In particular, the derivation is carried out under the assumption of simple random sampling. This is disturbing since in the case considered by the author, the units are selected with probability proportional to size. An additional assumption has also been made that μ_{4h}, the fourth moment about the mean is constant from stratum to stratum. There is no doubt that in spite of all the assumptions made, the Hansen-Hurwitz-Madow formula provides an approximation to the variance of the estimated variance. However, it is a question whether such an approximation can be used without justifying the assumptions made or adequate evidence concerning its reliability. As such the numerical results in respect of mean square error are of limited value.

viii) In evaluating the mean square error of the yearly variance estimate, it has been assumed that the variance of the yearly variance estimate is related to the variance of the monthly variance estimate in the same way as the variance of the yearly estimate is related to the variance of the monthly estimate. Some evidence supporting this assumption would greatly enhance the value of the results obtained.

I shall now discuss the other two papers which deal with presentation and analysis of sampling errors and design effects.

The paper by Krotki, Kish and Groves discusses the results based on eight fertility surveys from five different countries.

i) The authors have computed standard errors for about 40 variables spread over different classes and sub-classes. This is commendable and is likely to be appreciated by survey statisticians engaged in analysis of survey data and designing surveys.

ii) The authors have considered the important problem of presenting design parameters with a view to planning of future surveys. Among the

countries considered, some are highly developed while some are under-developed with high illiteracy rate. As such, the quality of data collected is likely to vary appreciably from one country to another. In the absence of any idea concerning the contribution of non-sampling errors, it is not clear whether the results from different surveys are at all comparable. Portability of such results is questionable. If results from countries with similar cultural background and level of development are available, they could perhaps be pooled together and such results may be useful for planning and designing of surveys of similar nature.

iii) The authors have proposed the use of intra-class correlation coefficients ρ and design effects as tools for designing future surveys. Since both the proposed parameters are functions of the type of stratification used, the selection procedures and sample sizes at different stages, they may not be portable unless the survey is to be repeated with only minor modifications. In fact ρ values obtained from different surveys may not be comparable unless the surveys are essentially designed in a similar manner.

iv) The section on summarizing sampling error results is informative and focuses attention on some important problems that arise and gives some guide lines as to how they can be tackled. The authors recognize the technical and analytical difficulties involved in combining and averaging results over different characteristics in a single survey. All the same, they recommend averaging irrespective of how the characteristics are related. Averaging over a group of related characteristics may be meaningful. If the characteristics are vastly different and ρ values are pooled and averaged, it is not clear what the average represents and whether it can at all be used in designing and planning future surveys.

v) It has been remarked that the ρ value for each characteristic and sub-class combination is subject to high variability but is quite stable when averaged over several sub-classes. This is only to be expected and cannot possibly justify averaging. It seems that by averaging we are giving away information concerning the variation in intra class correlation coefficient over different classes and characteristics. If averaging is considered essential, it may be desirable to give the range of values along with the average.

It is clear that under certain conditions, averaging would be desirable. There are several ways in which it can be carried out. This problem has been examined by Thomas Herzog with reference to 1973 Current Population Survey. He has considered several averaging methods including the one based on James-Stein estimator. The author concludes that the averaging method based on James-Stein estimator is the best among the lot. What is the basis of comparing the different averaging methods? It appears that the optimality criteria is subjective. It may be desirable to evolve suitable criteria consistent with practice and then compare the different methods.

REFERENCES

[1] Cochran, W. G. Sampling techniques, John Wiley and Sons, New York (1963).

[2] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. Sample Survey Methods and Theory Vol II, New York and London, Wiley Publication (1953).

[3] Hartley, H. O., Rao, J. N. K. and Grace Kiefer, Variance estimation with one unit per stratum, Journal of the American Statistical Association, 64, 841-851, (1969).

[4] Seth, G. R. On collapsing of strata, Journal of the Indian Society of Agricultural Statistics, 18, 1-3, (1966).

DISCUSSION

J.N.K. Rao, Carleton University, Ottawa, Canada

Section I of the paper by Grant Capps gives some theory for a generalized unequal probability sampling design which includes the usual with and without replacement designs as special cases. An interesting application to the Current Population Survey is given in Section II. The remaining two sections (III and IV) investigate a sample selection method which is a compromise between the one unit per stratum and the two units per stratum designs.

The generalized estimator of the population total Y considered in Section I is given by

$$\hat{Y} = \sum_{i=1}^{N} \frac{t_i}{E(t_i)} y_i , \qquad (1)$$

where $t_i (i = 1, \ldots, N)$ is the number of times the i-th population unit is included in a sample of fixed size $n (\Sigma t_i = n)$. Capps derived the variance of $\hat{Y}$ and two unbiased variance estimators from first principles. In this connection, it may be of interest to note that these results can be obtained simply from a general theorem (Rao and Vijayan [2]) which, in addition, gives the necessary form of nonnegative unbiased estimators of MSE. A general linear estimator of Y is given by

$$\hat{Y} = \sum_{i=1}^{N} d_{is} y_i \qquad (2)$$

where s denotes a sample selected according to a design $p(s)$, and the weights d_{is} in (2) are such that $d_{is} = 0$ if $i \notin s$. We have the following general theorem:

Theorem. Suppose the mean square of $\hat{Y}_d$ becomes zero when the ratios y_i/w_i are all equal for some constants $w_i (\neq 0)$. Then

(a) $MSE(\hat{Y}_d)$ reduces to

$$MSE(\hat{Y}_d) = - \sum_{i<j}^{N} \sum d_{ij} w_i w_j (z_i - z_j)^2 \qquad (3)$$

where $z_i = y_i/w_i$ and

$$d_{ij} = E(d_{is} - 1)(d_{js} - 1) = \Sigma_s p(s)(d_{is} - 1)(d_{js} - 1) ; \qquad (4)$$

(b) a nonnegative quadratic unbiased estimator of $MSE(\hat{Y}_d)$ is necessarily of the form

$$mse(\hat{Y}_d) = - \sum_{i<j} \sum d_{ij}(s) w_i w_j (z_i - z_j)^2 \qquad (5)$$

where $d_{ij}(s) = 0$ if s does not contain both units i and j, and

$$E(d_{ij}(s)) = \sum_{s \ni i,j} p(s) d_{ij}(s) = d_{ij} , \quad i < j . \qquad (6)$$

Equation (6) is the unbiasedness condition, and selected choices of $d_{ij}(s)$ satisfying (6) lead to unbiased estimators of $MSE(\hat{Y}_d)$. If $\hat{Y}_d$ is unbiased for Y as in the case of $\hat{Y}$, then $E(d_{is}) = 1$ and (4) reduces to

$$d_{ij} = E(d_{is} d_{js}) - 1 . \qquad (7)$$

We now illustrate the application of (3) - (7) to the estimator $\hat{Y}$ given by (1). The condition of our Theorem is satisfied with $w_i = E(t_i)$ and $y_i/w_i = c (\neq 0)$, since $\hat{Y}$ reduces to $c \Sigma t_i = cn$, a constant. Noting that $d_{is} = t_i/E(t_i)$ for $\hat{Y}$, we get from (4)

$$d_{ij} = cov(t_i, t_j)/\{E(t_i)E(t_j)\} , \qquad (8)$$

and (3) reduces to the formula (5) of Capps:

$$V(\hat{Y}) = - \sum_{i<j} \sum cov(t_i, t_j)(z_i - z_j)^2 . \qquad (9)$$

The choice

$$d_{ij}(s) = d_{ij} \frac{t_i t_j}{E(t_i t_j)} \qquad (10)$$

satisfies (6), and (5) reduces to

$$v(\hat{Y}) = - \sum_{i<j} \sum \frac{t_i t_j}{E(t_i t_j)} cov(t_i, t_j)(z_i - z_j)^2 \qquad (11)$$

which agrees with the formula (7) of Capps. The variance estimator (6) of Capps does not belong to the necessary class of nonnegative unbiased variance estimators, viz. (5).

The compromise scheme in Section III was obtained by choosing Scheme I (one unit per stratum design) with probability p and Scheme 2 (Durbin's scheme) with probability $1 - p$ $(0 \leq p \leq 1)$ and then selecting a sample of $n = 2$ units according to the chosen scheme. The variance formulae derived in Section III E (for the unconditional estimator $\hat{Y}_p$) and in Section III F (for the conditional estimator $\hat{Y}_c$) can be obtained simply from the general formulae (3) and (5) with the choice $d_{ij}(s) = d_{ij}/\pi_{ij}(p)$, $i < j \in s$. It also follows that (28) and (34) are the only possible nonnegative unbiased variance estimators for $\hat{Y}_p$ and $\hat{Y}_c$ respectively.

Fuller [1] has also proposed the compromise scheme, but confined himself to simple random sampling designs in which case $\hat{Y}_p$ and $\hat{Y}_c$ both reduce to $N\bar{y}$, where $\bar{y}$ is the sample mean. Fuller proposed an alternative method which appears preferable to the compromise scheme. The method is approximately as efficient as the one unit per stratum design and yet provides unbiased variance estimators. An extension of the alternative method to unequal probability sampling was also given.

REFERENCES

[1] Fuller, W.A. (1970). "Sampling with random stratum boundaries", Journal of the Royal Statistical Society, Ser. B, 32, 209-26.

[2] Rao, J.N.K. and Vijayan, K. (1977). "On estimating the variance in sampling with probability proportional to aggregate size", Journal of the American Statistical Association, 72, 579-84.

The paper by Isaki and Pinciaro gives useful empirical results on the relative performances of seven variance estimators for PPS systematic sampling. However, the study was confined to just one population, viz. mobile home dealers canvassed in the 1972 Census of Retail Trade. It would be useful if the study is extended to cover other real populations. Model-based investigations would also throw further light on the properties of the variance estimators. A model-based variance estimator proposed by Hartley [1] is not included in the study.

REFERENCES

[1] Hartley, H.O. (1966). "Systematic sampling with unequal probability and without replacement", Journal of the American Statistical Association, 61, 739-48.

SURVEYS: A PRACTICAL APPROACH

Nicholas J. Ciancio and John L. Stover, Statistical Reporting Service, USDA

I. INTRODUCTION

This paper will present to the student and those actually involved in the collection of survey statistics some of the problems encountered when developing, performing, and analyzing a survey. These aspects are usually omitted during the formal education of a statistician. Data used in textbooks or journal articles are usually presented without reference to the day-to-day problems of its collection, or are entirely hypothetical.

Recent literature such as Sudman [20], Sukhatme and Sukhatme [21], and Sudman [14] discuss some methodology of data collection, but for the most part the literature is entirely too theoretical concerning this topic. Examples of the latter are Cameron [2], Bryant, et al. [1], and Cornfield [6].

Therefore, the object of this paper will be to take the reader through a survey step by step. We will consider the survey as consisting of three main sections:

(a) Presurvey work,
(b) Collection of the data, and
(c) Data summary and estimation.

The discussion of the above concepts will be based on two surveys conducted in California. The first is a citrus weight study which is currently in the planning stages. The second is a survey of agricultural labor wage rates, which has been conducted by the Statistical Reporting Service of the USDA on a quarterly basis for the past decade. These surveys will be further discussed throughout this paper.

II. PRESURVEY WORK

Before the collection of data can be undertaken, adequate preparation must go into the initiation of the survey. This takes months of time. In the case of certain very large surveys, this can take several years. The first and foremost problem is to determine what information the results of the survey are to produce and who will supply this data. A target population must be determined, as well as the means by which the data are to be gathered.

Upon determining the target population, a reliable list of sampling units must be compiled. These lists are seldom complete, especially if there are a large number of participants. For the citrus weight study, the sampling unit is the packinghouse. The estimated number of packinghouses in California and Arizona is 317. In the farm labor survey, the employer of agricultural laborers was determined to be the sampling unit. There are approximately 13,000 farms and agricultural services in California employing such labor.

The method by which the data are to be collected will depend upon the results desired. For the citrus weight study it was determined to collect carton weights by size and grade so that an overall uniform weight by citrus variety can be determined. In the farm labor survey, wage rates, are desired for various types of farm workers (such as field, livestock, packinghouse, machine operators, etc.) and by method of pay (cash wages only, those given housing, those given room and board, etc.). Thus, employers are asked to give wages paid and hours worked for these different classes for a particular time period.

The problem of sample design type to be used will determine the technique of calculating sample size and allocation of the sample. For these topics, the student can find numerous texts that attack the problem. This is one of the most critical portions of the planning stage. If an incorrect sample size and allocation are used, the results of the survey will have little or no significance. If stratified random (or psuedo-random) sampling is used, estimates of the variances for each of the strata are needed. Such estimates are difficult to obtain for new surveys. Usually estimates of variances can be obtained by taking proportions of the range of possible values as determined by sampling distributions. Deming [8] gives specific ratios of the range, assuming one knows the range of values and a possible distribution.

A stratified random sample of cartons by variety was used for the citrus weight study. The sample size was calculated to be 23,400 cartons across variety by year. A Neyman allocation was used to distribute the sample into the strata. On the farm labor survey, a stratified psuedo-random sample of employers is used. Stratification is based on the peak number of farm laborers employed in a given year. The total sample is currently set at 1,645 employers in California. This sample size is adjusted periodically, based on strata variances computed from previous surveys.

The aforementioned problems of sample size and allocations cannot be taken too lightly. Much approximation goes into setting the sample size for the first time. However, for an ongoing survey, estimates can be made from previous year's data. At this point, the accuracy and the percentage confidence levels of the estimate are determined. Usually the statistician has to calculate sample sizes for various confidence levels and error rates so that a complete cost analysis can be drawn up.

The budget is of prime consideration. If costs of enumeration are high, then the sample size will be low. The largest portion of the budget will be wages paid to enumerators. Other costs to be considered include mileage paid to enumerators for the use of their cars, telephone expenses, training of enumerators and clerical staff, questionnaire design and printing, field equipment, computer expenses, and so forth. Other costs, depending on the type of survey, will also need to be included. Once the budget is exhausted, the survey is complete.

After the sample size and allocation are determined, the actual drawing of the sample from

the list occurs. These sampling units can then be plotted on a map to determine the number and location of enumerators and supervisors necessary to run the survey. Hiring of enumerators is a long and tedious task, particularly when there is no list of enumerators who have worked on similar previous surveys. Even if there is a list of enumerators, new people may have to be hired. It is recommended that some sort of screening device be used, preferably a test to determine the applicant's literacy, abilities to perform simple arithmetic calculations, read maps, and follow directions. Such a test has been found to be quite useful for hiring enumerators on surveys conducted by the Statistical Reporting Service. Inherent in this process is the determination of the qualifications necessary for the position. Overall the practice of hiring involves traveling to a certain area and staying in a set place for several days to conduct interviews. Carefully written advertisements placed in local newspapers and trade publications announcing the interview time and place are helpful. Other names of potential candidates can be obtained from currently employed enumerators, and from various county and municipal agents involved in a general way with the particular industry or segment of society to be studied. These last sources should not be relied upon exclusively if the evils of racism and sexism are to be avoided.

For the citrus weight study, it was determined that 70 packinghouses would suffice to obtain the necessary data using 7 enumerators to collect the data. On the farm labor survey it has been found that 65 enumerators and 5 supervisors are needed. When dealing with such numbers of people, who are really only intermittently employed on these types of surveys, it should be remembered that extra employees should be hired. Before any given survey, and frequently during the course of the survey, enumerators will quit (due to illness, death in the family, dislike of job, etc.) or will have to be terminated for incompetence. Their workload will then have to be reassigned to the remaining employees.

During this time frame, a questionnaire must be designed, printed, and tested in the field. It is necessary that the questionnaire be worded so that it is easily understood by the enumerators and respondents. Its execution must be well thought out so that it will provide precisely the data of interest. It is also very helpful to have it in a format so that key punchers can transcribe the data to computer cards. The questionnaire should not be too long if personal interviews are to be conducted. A long and thick questionnaire will increase the rate at which respondents refuse to cooperate as well as have a detrimental effect on the quality of the data obtained towards the end of the interview due to respondent fatigue.

Once the form is finalized, it is sent to the printer. It is recommended that galley proofs be obtained and checked for errors before the entire lot of questionnaires is printed. Also, a field test is useful to solve any problems that might crop up. Another item to watch for on Nationwide surveys is different meanings ascribed to certain words in various regions. Survey designers should take care in wording the questionnaire so that each question is understood uniformly by the respondents.

Preparation of computer software is necessary before the data collection is actually underway. This should include an edit system, summary program, and possible estimation procedures. These can either be prewritten software packages or self-written, tailored to individual needs. Sample data should be entered to check all possible errors and to insure that the summary, edit, and estimation work properly. Edit limits must be determined along with input formats for data entry.

Sets of instructions are needed for office procedures. These instructions should be exhaustive for preparation of enumerator supplies and their distribution, checking in questionnaires as they arrive, the clerical edit, statistical edits prior to key punching, and so forth. Procedure must also be documented for keeping track of enumerator assignments. The clerical staff must be schooled in exactly what is expected of them.

Similarly, instructions for enumerator procedures must be developed. They must also be self-sufficient and cover as many problem situations as possible. Included in these instructions are also procedures for filling out forms relating to the specific survey (such as time and mileage sheets, accident reports, overtime pay, and so on).

The equipment needed in the field should be ordered well in advance so that it arrives before the start of the survey, and there is sufficient time to allow for lost shipments and shipment of the wrong goods. Detailed inventories should be kept at all times of the location and quantity of supplies. In an on-going survey, care should be taken to insure that supplies are returned to the control of the statistician at the end of the survey.

A kit for each enumerator is made with enough equipment to carry the enumerator through the survey. These kits can be distributed at the training schools for enumerators. These sessions should be planned so that a reasonable number of people are in attendance. For large surveys, several schools may be necessary in different locations. These functions reiterate the material covered in the instruction manuals and disseminate administrative procedures. A classroom with proper lighting, space, and air-conditioning (or heating, depending on the season) is necessary. It is well worth the cost to rent such a room rather than make do with facilities that are not adequate. Enumerators need to be notified well in advance of the time and location of their respective schools. Should the school run more than one day, motel reservations will need to be made for those enumerators living out of the area in which the school is held. While conducting the school, practice interviews are made with most problems being covered. This gives the statistician the opportunity to "weed out" those employees who will not be able to perform adequately. The time frame of the survey should be discussed so that the enumerator will have an idea of what percentage of his work should be completed by specific dates during the course of the survey. At the end of the school, assignments and supply kits can be distributed.

III. DATA COLLECTION

During the training and hiring of enumerators, supervisory enumerators are also selected. They will have the important role of coordinating between the field and the statistician. The main assignment of the supervisor is to insure that quality data is collected. This can be done by quality control checks on the enumerators. This entails great effort and time, but is money well spent. A subsample of completed questionnaires can be made, and then verified, on a survey involving counts or objective measurement. This is more difficult on an interview type survey, as the respondent will be reluctant to give another interview. In this case, the supervisor can look over completed questionnaires for internal consistency and reasonableness. Any errors found can be corrected, and the enumerators informed of their mistakes. Should there be serious problems that cannot be corrected by the enumerator, the enumerator should be terminated, his assignment picked up and redistributed to employees who can do the job.

While the survey is in progress, the main problems are to insure that: the enumerators have sufficient supplies, they know their assignments and time frame, they actually collect the data in the field (as opposed to their living-rooms), and they submit this data to the statistician. Due to time limitations, it may be necessary to have the last few days' work shuttled to the statstician rather than relying on the postal service.

Once the data is in the office, it is checked in, clerical and statistical edits are made, and then the data is submitted to key punching. The backlog in editing, especially toward the end of the survey, may be a cause of concern.

Adjustments to assignments, and the possible hiring of new enumerators (done only as a last resort) are usually made during the course of the survey by the supervisors. Basically, the supervisor makes sure the data is collected correctly and returned on time.

At the end of the survey, it is necessary to collect all unused materials for reuse. Also, an evaluation of enumerator performance, based on supervisor reports and quality of data received in the office is very helpful.

Let us assume at this point of the survey that all the data that is going to be submitted has been edited and key punched. This does not mean all of the data has been submitted to the office. A certain percentage of the sample will not be accessible during the survey, and there will be respondents who refuse to cooperate. A decision must be made in how to handle these missing reports. If the estimate has to be submitted by a certain date, then a strict time-table must be adhered to. This means if data comes in after a certain point in time, it will not be used.

IV. DATA SUMMARY

A summary of the data can now be obtained since all the submitted data has been "cleaned" both clerically and statistically. The computer summary should include the raw data in tabular form by strata. Counts should be made on the number of completed and usable questionnaires. When more than one measurement is made on one unit, an analysis of variance table is helpful.

Expansions are calculated by dividing the sample size into the population size for each strata. To calculate an estimate, strata totals are multiplied by appropriate expansion factors. Estimates for missing reports must be included in strata totals.

If the summary is estimating a total that does not yield the answer needed directly, then some statistical technique is needed to approximate the final estimate. Some techniques are regressions, time series, chartings, and so on. Standard errors can be calculated by applying the formula applicable to the sampling technique employed.

V. CONCLUSION

The previous sections just touch upon some of the practical considerations that must be made in setting up and conducting a survey. As can be seen, setting up the survey consumes most of the time involved with the survey. Usually, time restrictions make the collection and analysis of the data move quickly. These sections of a survey pertain mainly to the types of surveys that the authors have encountered. Other surveys may have unique problems not discussed.

Briefly, a summary of main problems to watch for are:

(1) Budget -- Be very careful not to overrun the allocated funds. Give yourself sufficient financial room to operate.

(2) Hiring and training of quality enumerators -- This is essential to the reliability of the data. If poor data is the foundation of a project, then nothing but trash will be obtained, no matter how sophisticated the analysis.

(3) Time schedule -- Prepare well in advance of survey starting date. Whatever can go wrong probably will, so give yourself sufficient time to deal with various crises.

(4) List building -- Constantly revise and update the universe list. Especially helpful to enumerators are telephone numbers, physical addresses, (as opposed to post office box number), who to contact if interviewing a business, and who not to contact (John Smith, Sr. might give you the data, while John Smith, Jr. might drive you off with a shotgun).

(5) Sample size -- Revise according to the list. Proper estimating techniques should also be updated.

In this brief paper, the authors have presented some real world problems. Hopefully, this will help the reader who sets up and conducts surveys.

REFERENCES

[1] Bryant, E.C., Hartley, H.O., and Jessen, R.J., (1960). "Design and Estimation in Two-Way Stratification." J.A.S.A., 55, 105-123.

[2] Cameron, J.M. (1951). "The Use of Components of Variance in Preparing Schedules for Sampling of Baled Wool." Biometrics, 7, 83-96.

[3] Church, B.M. (1954). "Problems of Sample Allocation and Estimation in an Agricultural Survey." J.R.S.S. (B), 224-235.

[4] Cochran, W.G., (1963). "Sampling Techniques." John Wiley and Sons, Inc.

[5] Cornell, F.G., (1947). "A Stratified - Random Sample of a Small Finite Population." J.A.S.A., 42, 523-532.

[6] Cornfield, J., (1944). "On Samples from Finite Populations." J.A.S.A., 39, 236-239.

[7] Cox, D.R., (1952). "Estimation by Double Sampling." Brometrika, 39, 217-227.

[8] Deming, W.E., (1960). "Sample Design in Business Research." John Wiley and Sons, Inc.

[9] Fisher, R.A., (1973). "Statistical Methods for Research Workers." 14th ed., Hafner Publishing Company, New York.

[10] Hansen, M.H., Hurwity, W.N., and Madow, W.G., (1953). "Sample Survey Methods and Theory." John Wiley and Sons, Inc., Vol. 1.

[11] Jessen, R.J., (1942). "Statistical Investigation of a Sample Survey for Obtaining Farm Facts." Iowa Agr. Exp. Sta. Res. Bull., 304, 7-59.

[12] Jessen, R.J., and Houseman, E.E., (1944). "Statistical Investigations of Farm Sample Surveys in Iowa, Florida, and California." Iowa Agr. Exp. Sta. Res. Bull., 329, 265-338.

[13] Johnson, F.A., (1943). "A Statistical Study of Sampling Methods for Tree Nursery Inventories." Journal of Forestry, 41, 674-679.

[14] McVay, F.E., (1947). "Sampling Methods Applied to Estimating Numbers of Commercial Orchards In a Commercial Peach Area." J.A.S.A. 42, 533 540.

[15] Milne, A., (1959). "The Centric Septematic Area - Sample Treated as a Random Sample." Biometrics, 15, 270-297.

[16] Osborne, J.G., (1942). "Sampling Errors of Septematic and Random Surveys of Cover-Type Areas." J.A.S.A., 37, 256-264.

[17] Politz, A., and Simmons, W., (1949). "An Attempt to Get the 'Not at Homes' Into the Sample Without Callbacks." J.A.S.A., 9-31.

[18] Snedecor, G.W., and King, A.J., (1942). "Recent Developments in Sampling for Agricultural Statistics." J.A.S.A., 37, 95-102.

[19] Sudman, S., (1972). "On Sampling of Very Rare Human Populations." J.A.S.A., 67, 335-339.

[20] Sudman, S., (1976). "Applied Sampling." Academic Press, New York.

[21] Sukhatme, P.V., and Sukhatme, B.V., (1970). "Sampling Theory of Surveys with Applications." Iowa State University Press.

[22] Tukey, J.W., (1950). "Some Sampling Simplified." J.A.S.A., 45, 501-519.

[23] Yates, F., (1949). "Sampling Methods for Censuses and Surveys." Hafner Publishing Company, New York.

MEASURING INTERVIEWER PERFORMANCE IN MOCK INTERVIEWS

Anitra Rustemeyer, Bureau of the Census

INTRODUCTION

One aspect of the work undertaken by the Census Bureau's Committee to Evaluate Initial Training of Interviewers was to develop a means of evaluating interviewer skill during the conduct of an interview. A paper published in Britain in the early 1950's found that only 12% of all errors made by interviewers could be detected by review of completed interview materials turned in by an interviewer; the remaining 88% were "invisible" during later review of completed materials in that they resulted from altering the scope of questions, probing and prompting errors, and incorrect recording of information.[1]

In A Technique for Measuring Interviewer Performance, Charles Cannell summarized the results of work done in recent years at the University of Michigan's Survey Research Center to systematically and objectively measure interviewer on-job performance. SRC's method differs from that used in Britain and in the Census Bureau in that it makes use of tape recordings of live interviews conducted in respondents' homes; whereas, the Census Bureau and the British studies used mock interviews in which staff members role-played as respondents.

OVERVIEW OF STUDY DESIGN

For our attempt to develop a test of interviewer performance we selected three probability samples of Census Bureau interviewers. The three samples represented:

1. "New" interviewers, who had just completed their initial home study and classroom training for the Current Population Survey (CPS) and had no field experience with CPS (N=72);
2. "EOT" (end-of-trng.) interviewers, who had completed all phases of initial CPS training (including on-the-job training) and had completed two or three field interviewing assignments (N=39); and,
3. "Expr"(experienced) interviewers, who had completed all initial training and had more than three months of field experience on CPS (N=114).

Although interviewers were sampled according to their levels of experience, those tested do not represent the interviewer work force as a whole. The proportion of new interviewers selected was greater than the proportion of experienced interviewers.

Each interviewer selected for the study was asked to conduct either three or four interviews with a staff member.[2] The persons who role-played as respondents followed a script so that each interviewer was tested on nearly identical situations. Only if the interviewer asked incorrect questions was the respondent allowed to deviate from the script. All interviews were tape recorded. Coders listened to all of the tapes and coded the quality of asking questions, probing, and introducing and closing the interviews. They also reviewed the completed questionnaires and coded them for consistency with the tape recording. Care was taken during coder training and quality control operations to assure that only one error was assigned for each mistake, and that interviewers not be penalized for mistakes of the "respondent" (tester). Independent check coding maximized uniformity of coder decisions.

For each interviewer included in the study, actions were evaluated for the following aspects of interviewing:

- --Asking questions
- --Probing for additional information
- --Recording answers
- --Filling transcription items
- --Introductions and closings
- --Accuracy of labor force classification.[3]

The first five aspects of interviewing were viewed in three ways: (1) Proportion of actions of each type that were judged to be correct actions; (2) a "score" for each interviewer which was calculated to give relatively more weight to actions considered by the analyst to have greater impact on the quality of data; and (3) proportion of the five types of errors that were of each type (without regard to impact on quality of data).

To judge accuracy of labor force classification, the questionnaires filled by the interviewers were subjected to a coding process that duplicated as closely as possible the Census Bureau's computerized labor force classification system.

RESULTS

While interviewers were sampled, the test they took did not sample situations they meet at work. The scripts were designed to be graded from easy to somewhat difficult. Analysis of consistency in interviewers' scores according to script will indicate how much test results described here are affected by the difficulties presented in the scripts. The following results, therefore, should be viewed as provisional:

Proportion of Correct Actions: Written vs. Verbal. As can be seen in Table A, interviewers were correct more often in their written work than in the verbal part of their job. Written entries were of acceptable quality 94-97% of the time, while the way in which questions were asked was judged to be acceptable 84-89% of the time, and the way in which interviewers probed for additional information was judged to be acceptable a little over 80% of the time.

Types of Errors and Frequency of Errors. Table B identifies the nature of the seven scores which were computed for each interviewer and shows for each of the three interviewer groups the mean, range, and standard deviation of the scores. Statistically significant differences were found between experienced interviewers and new interviewers for three of the scores: experienced interviewers were significantly better at filling transcription items and entering notes required by the answers given by respondents; also, the score summarizing the quality of all written work showed that experienced interviewers were significantly

better than inexperienced interviewers in that aspect of their work.

Nearly one half of all errors were related to how well interviewers asked questions. New interviewers made significantly more errors than did the experienced ones. These findings can be seen in Table C. It is also interesting to note (from Tables B and C) the extent of individual variation in number of errors made and test scores.

Relationships among Test Scores and Other Information about Interviewers. In order to examine differences among the S scores, correlation coefficients were computed (some are presented in Tables D1 and D2). All of the relationships among scores S1-S6 for experienced and new interviewers are positive and statistically significant.

Because the testing procedure used in this study is relatively expensive to administer, it was important to determine whether it provided new information about interviewers or whether it was largely a duplicate of some other measurement already in use and/or available at lower cost. The relatively large number of small and insignificant correlation coefficients shown in Table E support the conclusion that this test of interviewer performance does not merely provide a different way to approximate an existing measurement.

Visible vs. Invisible Errors. In order to compare our findings with the British study referenced above, their classification scheme was applied (see Table F). In Britain, the most common type of error was "failure to probe," while in our study the most common error of experienced interviewers was to "alter the scope of the question"; the most common type of error made by the inexperienced interviewers was what the British called "invisible recording errors."

The new interviewers made the highest proportion of visible errors (33% of the errors classified in Table F); at the end of their training period 18% of the interviewers' errors were visible; finally, experienced interviewers had only 9% of their errors in the "visible" category.[4] While the experienced interviewers made fewer errors than did new ones, and apparently had learned to avoid errors in the "visible" category, they were much more likely to alter the scope of the question. Visible errors are relatively cheap and easy to correct because they can be detected by means of an office review; therefore, it is disconcerting that 91% of the errors made by experienced interviewers were "invisible."

Quality of Labor Force Classification. Table G summarizes labor force classification results. It shows that 36% of the experienced interviewers made one or more errors that would have prevented labor force classification or resulted in the wrong classification. Sixty-seven percent of the inexperienced interviewers made such errors, while 61% of those with two or three months of experience made errors that prevented labor force classification or resulted in misclassification.

When considering the findings shown in Table G (as well as those shown throughout this report), it is important to bear in mind that for this study the performance of interviewers was judged in an artificial setting. Whether interviewers performed better or worse in this setting than in the field is a matter of surmise, but this test is predicated on an assumption that there is a relationship between the way an interviewer behaves in the field and performs on the test. As noted at the outset, the situations protrayed in the scripts used in our test were chosen to present interviewers with a variety of test situations; they should not be interpreted as a representative sample of situations encountered during the conduct of the Current Population Survey. Within this restriction on the generalizability of these findings, it is worthwhile to note that this study does provide evidence that errors made while administering surveys can result in misclassification of respondents. Whether the percentage of persons misclassified is 6%, as we found in the situations we contrived for our test, or whether it is some other percent cannot be determined by this study.

Although the procedures used in this study are reasonably expensive to follow, we are hopeful that something similar to the test and coding procedures we developed can be implemented at the Bureau as a way of giving each interviewer and his supervisors objective feedback on how well the interviewer is performing in several aspects of his job.

FOOTNOTES

1. Reported in Harris, Muriel, "Interview-Research: Paper VI, The Grading of Interviewers: An Examination of Visible and Concealed Interviewer Error as Revealed by the Grading Tests, and Some Suggestions for Future Grading Procedure," M52, Documents Used During the Selection and Training of Social Survey Interviewers and Selected Papers on Interviewers and Interviewing, The Social Survey Division, Central Office of Information, Great Britain, May 1952.

2. Interviewers in groups 1 and 3 were tested on four scripts (A,B,C,D); those in group 2 were tested on three scripts (B,D,E). The same person role-played as the household respondent for all mock interviews administered by an interviewer; persons who played respondent were regional office supervisors or professional staff from the Bureau's Statistical Research Division.

3. This was included as a measurement of the effect of interviewer errors on the quality of final data.

4. The difference between new and experienced interviewers in percentage of visible errors is statistically significant, i.e., in a difference of proportions test an approximate Z value of 4 was obtained. EOT was not tested with the other groups because its sample was so small.

TABLE A DISTRIBUTION OF SELECTED BEHAVIORS, BY LEVEL OF INTERVIEWER EXPERIENCE

Type of Behavior	Level of Interviewer Experience					
	Experienced		End-of-Trng.		New	
	%Acceptable	% Unaccep.	%Acceptable	% Unaccep.	%Acceptable	% Unaccep.
I. VERBAL BEHAVIOR IN INITIAL ASKING OF QUESTIONS						
Total verbal behaviors coded for how questions were asked	22,255=100%		5,466=100%		14,089=100%	
	89.4	10.6	84.3	15.7	86.8	13.4
A. Asked questions exactly as worded	62.0		60.2		62.2	
B. Changed only verb tense	2.9		2.3		1.8	
C. Made minor modification more than verb tense	19.6		15.6		18.1	
D. Made correct use of verification	4.9		6.2		4.7	
E. Rephrased question & changed meaning; read answer categories when not permitted; or made improper use of verification in lieu of asking question		5.2		3.6		3.9
F. Asked a question which should have been skipped		2.1		3.2		3.5
G. Failed to ask a question which should have been asked		3.3		8.9		6.0
II. VERBAL BEHAVIOR AFTER INITIAL ASKING OF QUESTION (PROBING)						
Total number of probing behaviors coded	5,289=100%		1,202=100%		2,714=100%	
	80.7	19.4	86.0	14.1	80.3	19.6
A. Repeated questions correctly	5.6		7.6		7.1	
B. Made up non-directive probe	38.4		50.2		45.4	
C. Correctly repeated or summarized respondent	35.4		28.0		25.8	
D. Correctly confirmed frame of reference	1.3		.2		2.0	
E. Failed to probe when necessary		6.7		8.5		8.3
F. Probed directively		8.6		3.2		6.8
G. Incorrectly verified respondent's answer		2.2		1.2		2.2
H. Added to question incorrectly; repeated question or part of it incorrectly; confirmed incorrect frame of reference; or probed unnecessarily		1.9		1.2		2.3
III. WRITTEN BEHAVIORS	N=49,105		N=11,806		N=31,174	
	97.1	2.9	93.9	6.1	94.4	5.6
A. Number of items judged for quality of recording answers to survey questions	24,350=100%		6,069=100%		15,598=100%	
	95.5	4.5	91.2	8.8	94.9	5.1
1. Made correct entry	90.6		82.5		86.1	
2. Made entry consistent with verbal; but incorrect info was obtained due to interview verbal error	4.6		8.4		8.7	
3. Made entry consistent with verbal; but incorrect info was obtained due to respondent verbal error	.3		.3		.1	
4. Recorded info correctly; but it was not obtained in interview (usually means Ir guessed)		.5		1.7		.5
5. Entry or lack of entry was inconsistent with verbal (not used when (4) above applies)		3.9		6.2		4.5
6. Entry was in incorrect location; but intent was clear		.1		.3		.1
7. Omitted entry correctly		NA		.6		NA
B. Number of items judged for quality of filling transcription items	24,481=100%		5,576=100%		15,411=100%	
	98.8	1.3	97.8	2.2	94.2	5.8
1. Made a required entry correctly	98.8		97.8		94.2	
2. Made a required entry incorrectly		.8		.9		1.6
3. Failed to make a required entry		.5		1.3		4.2
C. Number of items judged for quality of entering required notes[1]	274=100%		161=100%		165=100%	
	88.7	11.3	59.0[2]	41.0[2]	70.9	29.1
1. Required note present and correct	88.7		59.0		70.9	
2. Required note not present		4.7		36.6		21.2
3. Required note present but not correct		6.6		4.4		7.9

TABLE A FOOTNOTES

[1]For the most part these were instances in which the interviewer marked an answer category which contained the instruction: "Specify" or "Specify in Notes".

[2]Comparison of EOT interviewers with the other two groups should not be made in this section as three-fourths of the "unacceptable" behaviors occurred in Script E, which was not used to test new and experienced interviewers.

TABLE B DESCRIPTION OF WEIGHTED SCORES ACHIEVED ON THE MIP TEST BY LEVEL OF INTERVIEWER EXPERIENCE

Type of Behavior Scored	Test Score[1]									t-Values for Expr. and New[2]
	Mean			Standard Deviation			Range			
	Expr	EOT	New	Expr	EOT	New	Expr	EOT	New	
S1. Asking Questions	574.9	473.2	524.0	155.9	137.9	149.4	178-905	178-811	164-860	1.931
S2. Probing	639.9	727.6	652.9	126.3	111.2	129.8	288-947	481-964	383-957	.033
S3. Written Entries	472.9	315.5	429.2	119.4	93.6	127.7	242-1000	119-540	183-982	2.169
S4. Recording Answers	434.4	286.1	407.8	121.4	95.4	129.3	214-1000	98-505	176-1000	1.288
S5. Filling Transcription Items	855.7	799.2	664.9	126.0	164.2	212.3	315-1000	364-1000	54-1000	7.089
S6. Entering Required Notes	790.1	353.5	524.1	369.7	234.8	440.1	000-1000[3]	000-1000[3]	000-1000[3]	4.592
S7. Introductions to and Closing of Interviews	399.5	361.6	433.9	201.7	128.9	147.3	141-1000	118-688	208-1000	-1.532

[1]The scores shown were computed on a scale of 0 to 1000, where 1000 is the best score possible. In forming the scores, some behaviors were given relatively more weight than others in order to reflect the opinion that all behaviors are not of equal importance. In computing the S scores, the weights were applied to the frequency counts and then the weighted count of acceptable behaviors was divided by the weighted count of all behaviors.

[2]In the test used, a positive t-Value means that the experienced interviewers were higher; a negative value means that new interviewers were higher. A value greater than 2 or less than -2 is statistically significant at the 5% level. EOT results were not tested with the other groups because of small sample size.

[3]This proportion is meaningless because it is often based on only 1 or 2 behaviors.

TABLE C SOME STATISTICS ABOUT THE NUMBER AND TYPE OF ERRORS MADE

Type of Error	Number of Errors Made									t-Values for Expr. and New[1]
	Mean			Stnd. Deviation			Range			
	Expr.	EOT	New	Expr.	EOT	New	Expr.	EOT	New	
Total Number of Errors Made	50.9	50.1	64.7	18.9	18.1	34.5	12-114	22-106	17-219	-2.722
E1. Asking	24.4	23.8	29.1	13.3	11.7	14.3	4-68	6-59	5-84	-1.823
E2. Probing	10.4	5.6	8.7	4.1	2.2	4.0	2-26	1-13	2-19	2.843
E3. Recording Answers	9.9	14.5	11.7	4.2	5.7	5.2	0-24	6-38	0-30	-2.307
E4. Transcription	2.6	3.1	12.4	3.0	4.0	22.3	0-21	0-20	0-138	-4.002
E5. All-Other Errors	3.6	3.1	2.7	2.1	1.3	1.6	0-11	1-7	1-9	4.169

[1]In the test used a positive t-Value means that the experienced interviewers were higher; a negative value means that new interviewers were higher. A value greater than 2 or less than -2 is statistically significant at the 5% level.

TABLE D1 CORRELATIONS AMONG TEST SCORES AND PERCENT OF ITEMS WITH PROBES FOR EXPERIENCED INTERVIEWERS

	S2	S3	S4	S5	S6	S7	Percent w/Probes
S1	.3750*	.5124*	.4953*	.2774*	.2613*	.0791	-.0561
S2		.3963*	.3858*	.2397*	.3001*	.1372*	.3791*
S3			.9960*	.3319*	.3358*	.0013	.1368
S4				.2624*	.3094*	-.0021	.1233
S5					.2413*	.0779	.1599
S6						.0505	.0483
S7							.0399

TABLE D2 CORRELATIONS AMONG TEST SCORES AND PERCENT OF ITEMS WITH PROBES FOR NEW INTERVIEWERS

	S2	S3	S4	S5	S6	S7	Percent w/Probes
S1	.2904*	.4934*	.4411*	.6346*	.3788*	.1819*	.1303
S2		.2447*	.2093*	.3570*	.3227*	-.0525	.2831*
S3			.9796*	.5459*	.2439*	.1248*	.0374
S4				.3938*	.2155*	.1235*	.0495
S5					.3083*	.0914	.0353
S6						-.1673*	.1808
S7							.1193

*Statistically significant at 5% level by Fisher's Z-statistic.

S1 -- Asking questions
S2 -- Probing
S3 -- Written entries (combination of S4,S5, & S6)
S4 -- Recording answers
S5 -- Filling transcription items
S6 -- Entering required notes
S7 -- Introductions and closings
Percent w/Probes--Percent of items on which probing was done

TABLE E CORRELATIONS (FOR EXPERIENCED INTERVIEWERS) BETWEEN TEST SCORES AND OTHER INFORMATION ABOUT INTERVIEWERS

	S1	S2	S3	S4	S5	S6	S7	D2	D3	D4	D5
D1	.0768	-.1536	.1614	.1654	-.0599	.1132	.0225	.0405	.0700	.0521	.0186
D2	-.3218*	-.2317	-.1672*	-.1692	-.0344	.0721	-.1725*		.0221	.2634*	-.0816
D3	-.2811*	-.2727	-.1264	-.1170	-.1455	-.3412	-.2193			.1217	.2490*
D4	.0469	.0563	.0581	.0560	.0071	.0514	.0605				.0056
D5	-.3073*	-.1904	-.0916	-.0874	-.0992	-.2810*	-.0445				

*Statistically significant at 5% level by Fisher's Z-test.

D1 -- Education
D2 -- Age
D3 -- Error Rate at time of test
D4 -- Number of minutes used to complete test
D5 -- Non-interview (of eligible households) Rate: weighted average for 3 months prior to test
S1 -- Asking questions
S2 -- Probing
S3 -- Written entries
S4 -- Recording answers
S5 -- Filling transcription items
S6 -- Entering required notes
S7 -- Introductions and closings

TABLE F DISTRIBUTION OF INTERVIEWER ERRORS BY TYPE AND LEVEL OF INTERVIEWER EXPERIENCE; COMPARED TO A BRITISH STUDY

TYPE OF ERROR	PERCENTAGE OF TOTAL ERRORS				
		Interviewers in Mock Interview Project (MIP)			
	Britain[2] (N=56)	Total (N=225)	Experienced (N=114)	End-of-Trng. (N=39)	New (N=72)
Total Errors[1]	1288=100%	7750=100%	3762=100%	1080=100%	2908=100%
Invisible Errors					
1. Failure to probe for additional information i.e. to find out if informant has anything further to add; and failure to probe sufficiently to establish criteria laid down in instructions or definitions or to clarify ambiguous answers.	34	9	9	9	8
2. Overprobing after it has become clear that informant has nothing further to add; or failure to recognize that they have all the information they require to classify.	7	1	1	1	1
3. Altering the scope of the question.	17	33	42	22	26
4. Prompting errors--failure to prompt when instructed, omission of items on prompt list, reading prompt list before all spontaneous information has been obtained.	1	8	9	7	6
5. Invisible recording errors. Any recording errors which could be discerned at the coding stage have been excluded from this category & dealt with the Category 6.	29	30	29	43	27
Visible Errors					
6. All errors discernible at the coding stage i.e. anything that appears to be an error in the light of other evidence on the schedule, omissions or inadequate information, items written in the wrong place and answers put under "others" when they fit a precode.	12	19	9	18	33

1. For the purpose of comparing the MIP results with the British study cited in Footnote 2, "total errors" is defined as it was in the British study. In the MIP the following additional types of errors were classified: incorrect selection of questions to be asked, asking questions out-of-order, and incorrect introduction to and closing of interviews.

2. Reported in Harris, Muriel, "Interviewer-Research: Paper VI, The Grading of Interviewers: An Examination of Visible and Concealed Interviewer Error as Revealed by the Grading Tests, and Some Suggestions for Future Grading Procedure," M.52, Documents Used During the Selection and Training of Social Survey Interviewers and Selected Papers on Interviewers and Interviewing, The Social Survey Division, Central Office of Information, Great Britain, May, 1952.

TABLE G SUMMARY OF LABOR FORCE CLASSIFICATION ERRORS OF INTERVIEWERS TESTED, BY DURATION OF EMPLOYMENT ON CPS

	All Interviewers (N=225)	Experienced (N=114)	End-of-Training (N=72)	New (N=39)
1. Percent of interviewers who made one or more errors affecting ESR classification	49.1	36.0	61.1	66.7
2. Number of persons portrayed in test scripts*	3008	1610	1008	390
3. Number of those on line 5 who were unclassifiable or misclassified	180	52	88	40
4. Line 3 as a percent of line 2	6.0%	3.2%	8.7%	10.3%
5. Mean number of unclassifiable and misclassified persons per interviewer	.80	.45	1.22	1.03

*This is the number of persons portrayed in each script, multiplied by the number of interviewers who were tested with the script.

COUNTING RULE BIAS IN HOUSEHOLD SURVEYS OF DEATHS

Monroe G. Sirken and Patricia N. Royston
National Center for Health Statistics
M. P. Bridges, Research Institute Triangle

INTRODUCTION

A test of the completeness of the death registration system in the United States has not been conducted because a suitable household sample survey has not been designed. Basically, a registration completeness test involves conducting a single time retrospective household sample survey to enumerate deaths and then matching the enumerated deaths with the file of registered death certificates. For some time, we have been working on a network household sample survey design for enumerating rare events which we feel has promise as an effective survey method for testing the death registration completeness [1]. The main innovation of the method relates to the counting rule. This rule defines the households that are eligible to report deaths in the household survey [2]. In the typical household sample survey, the de jure residence rule is used. According to this rule a decedent is eligible to be reported at only one address, namely the address of his former residence. Hereafter, we refer to this address as the key address. On the other hand, a network household sample survey design uses a counting rule that links deaths to networks of households of varying sizes which may or may not contain the households of the key addresses.

We have been investigating rules that link decedents to the households of their surviving relatives. The types of relatives covered by the counting rule must be specified carefully, however, to assure that the decedent is survived by at least one of them. Otherwise the decedent would have no chance of being enumerated in the survey. Counting rule bias is the proportion of decedents that is not linked to any households by the counting rule. In this paper, we present estimates of counting rule bias associated with several kinds of counting rules including (a) the de jure residence rule and (b) consanguine counting rules that link deaths to the households of specified surviving relatives, (c) rules that combine the features of (a) and (b). Also we present estimates of counting rule bias associated with geographic counting rules that circumscribe the area of the households linked to the death by (a), (b) and (c). For instance, one of the geographic rules limits eligibility to linked addresses within the county of the key address. Another limits eligibility to addresses in North Carolina.

DESIGN OF THE PILOT STUDY

Recently, we conducted a pilot study to investigate the error effects of the conventional counting rule and of consanguine and geographic counting rules on estimates of death registration completeness. In Stage 1 of this study we compiled a list of addresses of surviving relatives and key addresses for a sample of registered deaths. In Stage 2 we conducted interviews to see if the households at these addresses would report the deaths in the surveys. In Stage 3 we matched the deaths enumerated in the survey against the State file of registered death certificates.

The estimates of counting rule bias presented in this paper are based on the information collected in the first stage of the pilot study. Therefore, the design of the first stage is described in greater detail below.

A sample of about 1700 death records stratified by age and color was selected from death records on file in the State of North Carolina. Since the names and addresses of the death record informants are reported on the records, these people, who are generally close relatives of the decedents, were contacted by mail as soon as possible after the death was registered. They reported the names and addresses of specified surviving relatives, and the names of the occupants of the key households.

We limited the consanguine network to the relevant and closest relatives of the decedent. This varied depending on the age of the decedent. For decedents under 17 years of age, we obtained names and addresses of the decedent's mother (MO) and her parents (MP), and her siblings (MS). (For these decedents, the key address was defined as the address of the surviving mother.) For decedents aged 17-64, we obtained the names and addresses of the decedent's spouse (SP), siblings (SI), parents (PA) and children (CH), as well as the address of the key household (KH). For decedents 65 and over, we asked for the same names and addresses with the exception of parents.

FINDINGS

The estimates of counting rule bias are presented for four age groups in Tables 1-4. The stub of each of these tables lists the de jure residence rule and the various consanguine counting rules that were tested for the age group and the spread shows the three types of geographic counting rules. For each combination of consanguine and geographic counting rule separate estimates of counting rule bias are presented for (a) all deaths, (b) institutional deaths (decedents who were residents of long term institutions), and (c) noninstitutional deaths (decedents who were not residents of long term institutions). In the following discussion we illustrate our remarks with the findings for the age group 65-84 shown in Table 3.

The bias of the de jure residence rule is 21.8 percent. Actually, this represents the percentage of decedents who were residents of long term institutions and hence did not have a key household (KH). The bias of this rule is virtually zero for deaths that occurred outside of institutions since virtually all of them

formerly resided at a key address.

The bias of a counting rule decreases as the consanguine and geographic network expands. For instance, the bias of the rule that links decedents to surviving spouses (SP) residing in the key county is 56.6 percent. It decreases but slightly to 54 percent, when the geographic network is expanded to include spouses living anywhere in the United States. However, the bias decreases substantially when the consanguine network is expanded to include other types of relatives. If, in addition to the spouse (SP), decedents are linked to siblings (SI), or to siblings (SI) and children (CH), the biases decrease to 14.7 percent and 3.7 percent respectively. These figures imply that 54 percent of the decedents did not have a surviving spouse, 14.7 percent had neither a surviving spouse nor sibling, and 3.7 percent were not survived by a spouse, sibling or child. Viewed in this manner, the findings may be of substantive use to various social programs.

It is noteworthy that the counting rule bias was lowered only slightly, from 3.7 percent to 1.2 percent, by expanding the network to include key households (KH) in addition to the households of surviving spouses (SP), siblings (SI) and children (CH). This is one of the most important findings of the survey experiment. It reveals that use of the de jure residence rule is not mandatory to control counting rule bias. Certainly, it would be desirable to forego the conventional rule since the rule is difficult to implement and it is subject to large coverage bias [3]. For instance, more than 15 percent of the adult deaths in the pilot study represented people who were living alone when they died. In addition, 10 percent of the decedents formerly resided at a key address that was not occupied by any former members of his household within three months of his death. In total, a minimum of 25 percent of the key addresses were not occupied by a member of the decedent's former household by the date that the household survey was conducted, and consequently few of the households at these addresses reported any deaths in the survey.

In general, the biases of counting rules that link decedents to the households of their surviving close relatives increase with advancing age of the decedent. The survey using the counting rule that links decedents to their spouses (SP), siblings (SI) and children (CH) would fail to enumerate 10.7 percent of decedents over 85 years. By comparison, the rule that links decedents to the broadest network of close relatives is small for decedents of all age groups under 85 years. When decedents 65-84 years are linked to spouses (SP), siblings (SI) and children (CH), the bias is 3.7 percent. Moreover, the biases of consanguine counting rules are negligible for decedents in age groups under 65 years. The bias of the rule linking decedents under 17 years to households of their mothers (MO) is 1.1 percent and the bias is eliminated entirely if the decedents in this age group are also linked to the households of their mother's siblings (MS) and parents (MP). The bias is 4.2 percent if decedents 17-64 years are linked to spouses (SP) and siblings (SI) and it is only 1.7 percent if these decedents are also linked to parents (PA) and children (CH).

SUMMARY AND CONCLUSIONS

Counting rule bias in single retrospective household surveys that enumerate deaths varies by type of counting rule and by characteristics of decedent. The de jure residence rule fails to link institutional deaths to households where they would be enumerable in the survey. The seriousness of this problem increases with advancing age of the decedent. Thus, the counting rule bias of the de jure residence rule is 41.7 percent for decedents 85 years and older, 21.8 percent for decedents 65-84, 6.3 percent for decedents 17-64 and it is negligible for decedents under 17. Virtually all decedents under 65, whether or not they resided in an institution, are survived by close relatives of one type or another. Consequently, the bias of broad consanguine counting rules is negligible for these decedents. However, the bias of a broad consanguine rule is 3.7 percent and 10.7 percent respectively for age group 65-84 and 85 and over. If the de jure residence rule as well as a broad consanguine rule is adopted for these age groups the biases are reduced to 1.2 percent and 6.5 percent respectively.

It would be premature to evaluate counting rules entirely on the basis of counting rule bias [4]. Counting rules vary also in their effects on response bias and sampling errors. Error effects of these types were outside the scope of this paper. However, they will be the subject of a forthcoming paper.

REFERENCES

[1] Sirken, M.G., "Design of Household Sample Surveys to Test Death Registration Completeness," Demography, August 1973, Vol. 10, No. 3, pp. 469-478.

[2] _____, "The Counting Rule Strategy in Sample Surveys," Proceedings of the Social Statistics Section, American Statistical Association, (1974), pp. 119-123.

[3] _____, and Royston, P.N., "Under-reporting of Births and Deaths in Household Surveys of Population Change," Proceedings of the Social Statistics Section, American Statistical Association, (1973), pp. 412-415.

[4] _____, and _____, "Design Effects in Retrospective Mortality Surveys," Proceedings of the Social Statistics Section, American Statistical Association, (1976), pp. 773-777.

Table 1. Counting Rule Bias (in percent) by Counting Rule and Place of Residence at Death: Decedents Under 17 Years

Consanguine Counting Rule*	Geographic Counting Rule								
	United States			North Carolina			Key County		
	All Deaths	Residence at Death		All Deaths	Residence at Death		All Deaths	Residence at Death	
		Institution	Other		Institution	Other		Institution	Other
MO	1.1	--	--	5.1	--	--	5.1	--	--
MS	6.9	--	--	27.2	--	--	40.8	--	--
MP	5.5	--	--	25.4	--	--	40.6	--	--
MO+MS	0.0	--	--	2.5	--	--	5.1	--	--
MO+MP	0.4	--	--	3.0	--	--	5.1	--	--
MS+MP	0.9	--	--	19.2	--	--	32.3	--	--
MO+MS+MP	0.0	--	--	2.1	--	--	5.1	--	--

-- Not applicable
* See glossary in appendix for explanation of abbreviations.

Table 2. Counting Rule Bias (in percent) by Counting Rule and Place of Residence at Death: Decedents 65-84 years

Consanguine Counting Rule*	Geographic Counting Rule								
	United States			North Carolina			Key County		
	All Deaths	Residence at Death		All Deaths	Residence at Death		All Deaths	Residence at Death	
		Institution	Other		Institution	Other		Institution	Other
SP	40.0	74.2	37.7	43.6	77.5	41.3	46.1	88.4	43.2
SI	11.8	14.5	11.6	28.6	23.3	28.9	44.1	52.7	43.5
PA	56.3	84.1	54.4	64.9	84.1	63.6	73.5	100.0	71.7
CH	28.5	54.4	26.7	36.6	62.8	34.9	43.3	73.7	41.2
SP+SI	4.2	14.5	3.5	8.5	17.8	7.8	17.0	52.7	14.6
SP+PA	23.6	63.7	20.9	28.1	67.0	25.5	32.7	88.4	29.0
SP+CH	20.1	48.3	18.2	25.1	56.7	23.0	29.2	67.6	26.6
SI+PA	7.0	10.9	6.7	22.8	19.6	23.0	38.1	52.7	37.1
SI+CH	4.3	14.5	3.6	10.0	14.5	9.7	19.6	43.5	18.0
PA+CH	15.9	46.9	13.8	24.0	55.3	21.9	32.5	73.7	29.7
SP+SI+PA	2.6	10.9	2.1	6.2	14.2	5.6	15.1	52.7	12.5
SP+SI+CH	2.6	14.5	1.8	5.1	14.5	4.5	11.2	43.5	9.0
SP+PA+CH	10.6	40.7	8.5	14.7	49.1	12.4	20.3	67.6	17.1
SI+PA+CH	3.0	10.9	2.5	8.3	10.9	8.1	17.6	43.5	15.8
SP+SI+PA+CH	1.7	10.9	1.1	3.6	10.9	3.1	10.1	43.5	7.8
KH	6.3	100.0	0.0	6.5	100.0	0.0	6.5	100.0	0.0
SP+KH	4.7	74.2	0.0	4.9	77.5	0.0	5.8	88.4	0.0
SI+KH	0.9	14.5	0.0	1.7	23.3	0.0	3.5	52.7	0.0
PA+KH	5.3	84.1	0.0	5.5	84.1	0.0	6.5	100.0	0.0
CH+KH	3.4	54.4	0.0	4.2	62.8	0.0	4.9	73.7	0.0
SP+SI+KH	0.9	14.5	0.0	1.1	17.8	0.0	3.5	52.7	0.0
SP+PA+KH	4.0	63.7	0.0	4.2	67.0	0.0	5.8	88.4	0.0
SP+CH+KH	3.1	48.3	0.0	3.6	56.7	0.0	4.5	67.6	0.0
SI+PA+KH	0.7	10.9	0.0	1.4	19.6	0.0	3.5	52.7	0.0
SI+CH+KH	0.9	14.5	0.0	1.1	14.5	0.0	2.9	43.5	0.0
PA+CH+KH	3.0	46.9	0.0	3.7	55.3	0.0	4.9	73.7	0.0
SP+SI+PA+KH	0.7	10.9	0.0	0.9	14.2	0.0	3.5	52.7	0.0
SP+SI+CH+KH	0.9	14.5	0.0	0.9	14.5	0.0	2.9	43.5	0.0
SP+PA+CH+KH	2.6	40.7	0.0	3.1	49.1	0.0	4.5	67.6	0.0
SI+PA+CH+KH	0.7	10.9	0.0	0.9	10.9	0.0	2.9	43.5	0.0
SP+SI+PA+CH+KH	0.7	10.9	0.0	0.7	10.9	0.0	2.9	43.5	0.0

*See glossary in appendix for explanation of abbreviations.

Table 3. Counting Rule Bias (in percent) by Counting Rule and Place of Residence at Death: Decedents 65-84 years

Consanguing Counting Rule*	Geographic Counting Rule								
	United States			North Carolina			Key County		
	All Deaths	Residence at Death		All Deaths	Residence at Death		All Deaths	Residence at Death	
		Institution	Other		Institution	Other		Institution	Other
SP	54.0	85.1	45.3	55.4	85.7	47.0	56.6	88.8	47.7
SI	22.3	26.9	21.1	38.9	45.5	37.1	59.3	72.8	55.5
CH	24.2	31.4	22.2	30.6	36.6	28.9	37.3	54.7	32.4
SP+SI	14.7	21.7	12.7	22.9	34.9	19.6	35.9	63.9	28.1
SP+CH	16.0	29.9	12.2	19.3	35.1	14.9	25.0	54.2	16.9
SI+CH	5.7	5.3	5.8	12.3	11.1	12.6	23.2	36.2	19.6
SP+SI+CH	3.7	5.3	3.2	7.9	11.1	7.0	16.1	35.7	10.7
KH	21.8	100.0	0.0	22.1	100.0	0.0	22.1	100.0	0.0
SP+KH	18.5	85.1	0.0	19.0	85.7	0.0	19.7	88.8	0.0
SI+KH	5.8	26.9	0.0	10.3	45.5	0.0	16.2	72.8	0.0
CH+KH	6.8	31.4	0.0	8.0	36.6	0.0	12.3	54.7	0.0
SP+SI+KH	4.7	21.7	0.0	8.0	34.9	0.0	14.3	63.9	0.0
SP+CH+KH	6.5	29.9	0.0	7.6	35.1	0.0	12.1	54.2	0.0
SI+CH+KH	1.2	5.3	0.0	2.4	11.1	0.0	8.2	36.2	0.0
SP+SI+CH+KH	1.2	5.3	0.0	2.4	11.1	0.0	8.1	35.7	0.0

*See glossary in appendix for explanation of abbreviations.

Table 4. Counting Rule Bias (in percent) by Counting Rule and Place of Residence at Death: Decedents 85 years and over

Consanguine Counting Rule*	Geographic Counting Rule								
	United States			North Carolina			Key County		
	All Deaths	Residence at Death		All Deaths	Residence at Death		All Deaths	Residence at Death	
		Institution	Other		Institution	Other		Institution	Other
SP	84.0	95.4	75.9	84.1	95.4	76.1	84.1	95.4	76.1
SI	56.5	58.5	55.0	64.8	63.9	65.4	79.9	82.9	77.8
CH	23.6	31.3	18.2	25.2	32.0	20.4	35.9	52.2	24.4
SP+SI	48.2	57.6	41.4	54.8	62.9	49.0	67.1	80.6	57.5
SP+CH	20.8	29.1	14.9	21.4	29.9	15.4	31.2	50.0	17.7
SI+CH	12.9	16.6	10.2	15.4	19.2	12.6	28.8	43.8	18.0
SP+SI+CH	10.7	15.7	7.2	12.7	18.3	8.6	24.6	42.9	11.6
KH	41.7	100.0	0.0	42.1	100.0	0.0	42.1	100.0	0.0
SP+KH	39.8	95.4	0.0	40.2	95.4	0.0	40.2	95.4	0.0
SI+KH	24.4	58.5	0.0	27.1	63.9	0.0	35.0	82.9	0.0
CH+KH	13.0	31.3	0.0	13.3	32.0	0.0	22.2	52.2	0.0
SP+SI+KH	24.0	57.6	0.0	26.7	62.9	0.0	34.1	80.6	0.0
SP+CH+KH	12.1	29.1	0.0	12.4	29.9	0.0	21.3	50.0	0.0
SI+CH+KH	6.9	16.6	0.0	8.0	19.2	0.0	18.7	43.8	0.0
SP+SI+CH+KH	6.5	15.7	0.0	7.6	18.3	0.0	18.3	42.9	0.0

*See glossary in appendix for explanation of abbreviations.

APPENDIX: Glossary of Terms

Network survey: The events being enumerated are linked to networks of households.

Counting rule: Defines the network of households which are eligible to report events in a survey.

Counting rule bias: The fraction of events that are not linked to any households by the counting rule.

Key address:

Decedents over 16 years: Address of the noninstitutional decedent at the time of death.

Decedents under 17 years: Address of the decedent's mother at the time of the survey.

Counting rule abbreviations:

MO....Mother
MS....Maternal siblings
MP....Maternal grandparents
SP....Spouse
SI....Siblings
PA....Parents
CH....Children
KH....Key household
(Household occupying the key address)

Consanguine counting rule: A rule that links decedents to the households of surviving relatives.

Geographic counting rule: A rule that circumscribes the area within which the eligible households must be located.

EFFECT OF ETHNICITY OF SIGNATURE IN MAIL SURVEYS

Hershey H. Friedman
Department of Administrative Sciences
Montclair State College

Robert B. Fireworker*
Division of Business
St. John's University

Abstract

The current study was a follow-up to a previous one conducted by the senior author which showed that Hispanic and Jewish names did not affect the rate of response or content of a mail questionnaire sent to travel agents. The present research showed that an apparently Black name was as effective as a "WASP" name in eliciting returns to a mail survey. The content of the questionnaire, an "attitudes towards Jews" scale, was not influenced by the ethnicity of the sender's signature.

Using a sample of travel agents, Friedman and Goldstein (1975) found no difference in the return rate of, or responses to, a mail questionnaire signed by a Jewish, Hispanic, or ethnically unidentifiable name. The purpose of the current study were: (1) to determine whether a Black name would affect the rate of response or content of a mail survey, and (2) to determine whether ego-involving questions would produce results similar to the original study.

Method

On the basis of a pretest using 73 New Jersey residents--36 males and 37 females--the authors decided to use "Leroy Jefferson" as the Black name. The other name used in the study was "John Carter III." It was identified as a "WASP" name by 63 subjects in the pretest, while 6 subjects could not associate the name with any particular ethnic group.

The sample for the study consisted of 200 people randomly selected from a Northern New Jersey directory. One hundred subjects were sent the questionnaire with "Leroy Jefferson" identified as the sender. The other 100 subjects were sent the same questionnaire with "John Carter III" identified as the sender. In the cover letter, the sender identified himself as a graduate student conducting research on attitudes concerning race and religion at a Northern New Jersey college.

Subjects were sent the short form of the Levinson and Sanford (1944) anti-Semitism scale. The original scale consisted of 52 six point Likert-type statements. The short form consisted of ten statements selected from the original 52, both on a statistical and theoretical basis. Reliabilities of .89 to .94 have been reported for the short form of the A-S scale (Robinson & Shaver, 1973, pp. 371-378). The scale consists of ten statements, all of which express unfavorable attitudes towards Jews. For instance, "I can hardly imagine myself marrying a Jew." Subjects checked the amount of agreement as one of the following: strong agreement, moderate agreement, slight agreement, slight disagreement, moderate disagreement, strong disagreement. The responses were scored +3, +2, +1, -1, -2, -3, respectively. Thus the possible range of scores was from +30, for a strongly anti-Semitic individual, to -30, for a strongly pro-Semitic individual. It was believed that subjects responding to "Leroy Jefferson" would exhibit less racist tendencies than those responding to "John Carter III." An "attitude towards Jews" scale was used, rather than an "attitude towards Blacks" scale, in order to disguise the true purpose of the study. It was felt the use of an "attitude towards Blacks" scale with "Leroy Jefferson: as the sender might cause subjects to be suspicious.

Results and Discussion

Twenty-three of "Leroy Jefferson"s questionnaires, and 27 of "John Carter III's questionnaires, were returned. The chi-square value was not significant, $\chi^2_{(1)} = .43$. The mean scores on the anti-Semitism scale were -14.6 for the "Jefferson" group (variance = 98.0), and -14.6 for the "Carter" group (variance = 81.8). The t-value was obviously not significant at $t_{(48\ d.f.)}=0.0$.

This study provides further evidence for Friedman and Goldstein's (1975) contention that the warning of many researchers not to use ethnically identifiable names may not be valid. In addition, the use of ego-involving statements in the current study did not alter the results.

The limitations of the study are, of course, the small sample size used and the low response rate. Also, the possibility exists that bigots chose not to respond to the questionnaire, regardless of the name of the sender.

The current study was replicated by the author, using a similar methodology. The rates of response and the mean scores of the responses to an attitude toward Blacks scale did not differ, whether a "Black" name or ethnically neutral name was used. Thus, it seems, researchers may not have to disguise their names when sending out questionnaires in order to make them appear ethnically neutral.

Footnote

*The authors wish to acknowledge the invaluable assistance of Barbara Poda and Garrison D. Miller without which this study could not have been completed.

References

Friedman, H.H. and L. Goldstein. Effect of ethnicity of signature on the rate of return and content of a mail questionnaire. Journal of Applied Psychology, 1975, 60, 770-771.

Levinson, D.J. and R.N. Sanford. A scale for the measurement of anti-Semitism. Journal of Psychology, 1944, 17, 339-370.

Robinson, J.P. and P.R. Shaver. Measures of Social Psychological Attitudes. Ann Arbor, Michigan: University of Michigan Institute for Social Research, 1973.

CODING OF VERBAL INFORMATION

Lars Lyberg, National Central Bureau of Statistics, Sweden

1 CODING

As the term 'coding' has several meanings in various contexts, we give a short review of its use here.

Coding is a major operation of such statistical studies as, for instance, a census of population. It is assumed that every element E_1, E_2, ..., E_N in the population belongs to one and only one of, say k, categories. Usually written information about the element is obtained on schedules. For the purpose of data processing such written information must practically always be converted to numbers ("codes"). This act of converting is called coding although a better word might be 'classification'.

2 THE ERROR PROBLEM

There is ample evidence that the coding operation may be rather susceptible to errors: elements are not assigned into proper categories. As a consequence, there is need for control. The error rate is in fact substantial in many statistical studies. Gross errors of 10-25% when coding multi-digit difficult variables such as occupation and industry are not unusual. The solutions to the error problem have so far mainly consisted of methods for intense training and education of clerks along with the use of more or less efficient verification systems.

However, there are new approaches which will be touched upon in this paper. One example is automatic coding. Despite large gross error rates the net error rate could sometimes be very small. In the 1970 U S census coding of industry and occupation, gross error rates of 9 and 13 percent respectively were estimated. In Jabine and Tepping (1973) it is shown that this error rate results only in a relatively small contribution to the total mean square error for the two variables. Obviously the effect of coders and their error is small in some studies but in others the effects could be alarming. In the U S studies the small effects were obtained in a quality controlled material. Many surveys have no such program and if they have it could be a rather inefficient one. But the problem becomes acute having the forthcoming era of data bases in mind. Suppose we want to study subpopulations such as "people in retail trade". A gross error rate of 10% could be a very serious drawback in this situation. The coding errors result in over- and undercoverage.

3 SOME STUDIES OF ERROR RATES AT THE NATIONAL CENTRAL BUREAU OF STATISTICS, SWEDEN (SCB)

3.1 CODING IN LABOR FORCE SURVEYS

One early study described in Olofsson (1965) treats the "variability in occupation and industry data in Labor Force Surveys". There it is shown that coding errors are seriously affecting the estimates of changes such as the flow between different occupation and industry categories. The main result of the study was that only 40% of the changes in major occupation categories were real changes. The corresponding estimate for industry was 46%. The rest was due to coding errors. As a consequence an exaggerated picture of the mobility in the labor market is created. In fact the coding errors lead to overestimations of 100-200% for some categories.

3.2 CODING ERRORS IN THE 1965 SWEDISH CENSUS OF POPULATION

An evaluation study of coding errors was carried out in connection with the 1965 Swedish census of population. The study is described in Lyberg and Dalenius (1968). The modest prime objective of the study was to illuminate, in a concrete fashion, the performance of the dependent verification used in the census. As a by-product an evaluation of the coding was obtained. Here some selected results are given.

From a population of census material comprising about 70 percent of the 1965 population a two-stage sample of verified census schedules was selected. The population was partitioned into four strata subsequentally resulting in four subsamples. The evaluation study contained the following four variables:

(1) Relationship to head of household
(2) Employment
(3) Occupational status
(4) Industry

The codes used for the variables 1-3 were one-digit-codes; the code used for 'industry' was a three-digit-code.

The samples were coded by a team of three experimental coders. Each coded independently of the others. After that the codes were matched and three cases could occur. First, all three coders could agree; we call that case 3-0. Secondly two could agree but not the third; we call that case 2-1. Finally no two coders agree; we call that case 1-1-1. Apparantly in the first and second cases we are able to define a majority code. We used that code as an evaluation code. When the third case appeared we let a 'super-expert' decide an evaluation code.

Let us give the results for the three-digit variable (4) (industry). Table 1 a-b. A comparison between dependent and independent verification: the majority code M_4 is compared to P_4 and V_4. P_4 means production coder and V_4 means verifier.

Table 1 a

Experimental coder combinations	V_4 agrees with 3	2	1	0	Super expert cases	Total
3-0	451	-	-	24	-	475
2-1	-	44	23	6	-	73
1-1-1	-	-	-	-	5	5
						553

Table 1 b

Experimental coder combinations	P_4 agrees with 3	2	1	0	Super expert cases	Total
3-0	427	-	-	48	-	475
2-1	-	41	24	8	-	73
1-1-1	-	-	-	-	5	5
						553

This is a difficult variable to code. The experimental coders agree only in 475 of the 553 cases (86 percent). The error rate for the production coder is 80/548 or 14,6% and the associate figure for the verifier is 53/548 or 9,7%. As could be seen from the tables the dependent verification system reduces the error rate but the reduction is rather modest. In fact the tables illustrate the well known experience that dependent verification is rather ineffective. Especially the reduction is very small among the 2-1 cases. A possible explanation is that those cases are hard to code and that the coder has a tendency to let an assigned code remain unchanged. But even when we are dealing with the 3-0 cases only a 50% reduction in error rate is registered.

The code for multi-digit variables is often built on the principle of chineese boxes; i.e. the first digit stands for a major classification, the second digit for a classification within this major group etc. This is the case for the industry variable. Usually an error on the first digit is more serious than an error on the second and so on. We have studied the distribution of errors on the different digits for the industry varaible. Let us consider the deviations in the tables above.

Table 2 a-b. Frequency of deviations between M_4 and P_4, and M_4 and V_4 on first, second, and third digit level.

Table 2 a

	P_4			
Digit	First digit	Second digit	Third digit	Total
Deviation cases	41	17	22	80

Table 2 b

	V_4			
Digit	First digit	Second digit	Third digit	Total
Deviation cases	28	9	16	53

As could be seen from the tables most errors are serious; i.e. the error occurs already on the first digit (major group classification).

3.3 CODING ERRORS IN THE 1970 SWEDISH CENSUS OF POPULATION

In the 1970 census of population some improvements concerning the coding quality control program were carried out. For instance, about one third of the schedules was controlled by means of independent verification. However, one third was controlled by dependent verification and for the rest the quality measures were only estimated. So there was a need for an evaluation study. The primary goal for this study was to estimate the coding error rate after verification. A nationwide sample of 7 000 individuals was selected. The population was separated in three different strata reflecting the fact that three different control programs had been used.

Stratum 1: Dependent verification on a 100 percent basis

Stratum 2: Independent verification on a 10 percent sampling basis using an acceptance sampling plan

Stratum 3: Independent verification on a 10 percent sampling basis without using an acceptance sampling plan.

A pool of expert coders was used to generate a set of 'true' evaluation codes for each schedule in the sample. These codes were compared with the production codes after verification and this led to estimates of error rates for the different variables on economic activity. These variables were

(1) Relationship to head of household
(2) Type of activity
(3) Occupation
(4) Status
(5) Industry
(6) Kind of employment
(7) Way of travel to place of work
(8) Amount of hours at work

Variable (3) was a three-digit one and variable (5) was a four-digit one. The rest were one-digit ones.

In table 3 estimates of error rates for these variables are given.

Table 3 Estimated error frequency (%)

Variable	Percent error rate Stratum 1	2	3	Total population
(1)	4.5	3.8	5.1	4.3
(2)	4.4	5.3	4.0	4.7
(3)	12.6	12.7	16.5	13.5
(4)	4.2	3.1	3.8	3.7
(5)	8.8	9.9	11.6	9.9
(6)	9.5	10.7	5.4	8.9
(7)	11.0	11.3	12.6	11.5
(8)	4.0	4.2	5.4	4.4

The table shows that the multi-digit variables are difficult to code but even the one-digit variables are erronously classified to a relatively large extent. One reason could be that the coding situation is too complex for one coder, i.e. each coder has too many variables to manage. The errors on occupation and industry have the same pattern as has been shown in earlier studies. Most errors occur already on major group classification. Thus a coding error on these variables is often a serious error.

We also calculated the within expert coder variability WV defined as

$$WV = \frac{x}{n}$$

where n is the number of coded individuals in the experiment and where x is the number of unequally coded individuals in two independent trials.

For the five experts in the expert pool the following results were obtained.

Table 4 Within expert coder variability (%)

Variable	Expert A	B	C	D	E
(1)	0.7	1.2	2.4	1.1	0.8
(2)	1.2	2.1	3.0	1.5	1.8
(3)	8.0	10.6	10.9	9.2	7.1
(4)	2.4	0.9	1.8	1.1	1.9
(5)	3.7	8.8	11.6	6.9	5.4
(6)	0.8	2.7	6.0	1.4	2.9
(7)	1.3	1.5	2.1	1.8	2.5
(8)	1.6	3.2	3.9	2.7	2.1

The variability is substantial although these coders have worked for several years with this kind of coding.

4 ALTERNATIVE APPROACHES TO ERROR CONTROL

The control of coding operations could be carried out in many different ways. Some approaches are

- evaluation of coding results
- training and education of clerks
- the use of verification systems
- improving dictionaries and clerk manuals
- using automatic coding.

A total coding quality control system involves more than one of these approaches.

Evaluation of classification results is the basis for dimensioning the quality control efforts. We have already given examples of different evaluation studies. The results of such studies give hints concerning the size of the necessary quality control program.

Evaluation systems are based upon the existence of 'true' codes which are generated by means of more skilled clerks or expert coders. These true codes are compared to those assigned by the production coders and an estimate of production coding gross error rate could be calculated. Evaluation studies are, for instance, found in Fasteau et al (1962), Fasteau et al (1964), Minton (1969), Jabine and Tepping (1973) and U S Bureau of the Census (1972).

The training and education of clerks is indeed valuable since the error rate curve often decreases with time. If it is possible to 'cut' error rates at the beginning of a coding operation one will probably get a more acceptable average outgoing quality.

The literature covering this field is not especially extensive. However, the subject is discussed in Minton (1969) and in Dalenius and Frank (1968). In the latter the idea about using master sets is presented. A master set is a set of elements for which the correct classification is known. Such a set could be used during the training period and as a device for controlling the production process.

The use of verification systems is important to keep up the aimed at quality level. However, the systems could sometimes be rather inefficient, i.e. errors of type I and type II could occur.

The impact of these errors on single sampling plans is discussed in Minton (1972). The flow of coders between total and sampling controls is another problem. The flow must be regulated by means of some prespecified criterion. In Cook (1961) a special point system is given, where each coder receives a point for each erroneous coding. In Minton (1970) some other decision rules for administrative applications of quality control are discussed.

There are two main schemes for verification of coding. These are called dependent and independent verification. Dependent means that the verifier has access to the code assigned by the production coder. Independent means that the verifier has no such access and that the decision upon outgoing code must be based on different rules such as majority or modal rules. Within these schemes several realistic sub-schemes could be defined. The schemes could be used on a total or on a sampling basis. We have seen that dependent

verification could be rather ineffective. Many errors are not corrected. On the other hand the superior independent systems are more costly. Dependent and independent verification is dealt with in Lyberg (1967), Lyberg (1969) and Minton (1969).

Obviously many of the coding errors do not depend on the ability of clerks. Often the dictionaries and the clerk manuals are insufficient and cause a great variability in the coding process.

It is possible to use automatic coding in order to master the variability problem and to speed up the whole operation. Verbal discriptions of the variable under consideration are fed into a computer, a built-in dictionary is consulted and codes are assigned by the computer.

5 AUTOMATIC CODING

Automatic coding might be a complement to manual coding. The method has its strength in speeding up the entire operation but it could also be an instrument for reducing the coding variability. The method is described in O'Reagan (1972) and the main components are the following.

The verbal information for an element is transferred to a punchcard or a magnetic tape. Then the information is fed into a computer where a dictionary is stored. The information is matched against the descriptions in the dictionary. If match occurs the element is coded. Otherwise the element is sorted out and coded manually. The system for automatic coding must also contain continuous evaluation.

5.1 THE COMPUTER-STORED DICTIONARY

The dictionary should replace the coding instructions used in manual coding. Thus the construction of such a dictionary is very important. The construction work could be done manually in simple applications, but when dealing with multidigit variables we must have support from the computer. There are several steps in this work, for instance:

A Choise of a basic material
B Sampling a basic file from the basic material
C Expert coding of the basic file
D Establish inclusion criterias
E Construction of preliminary dictionary
F Testing and making complementary additions and reductions in the dictionary.

The basic material should ideally consist of the material to be coded. If you want to apply automatic coding in the 1980 census the dictionary should be based on descriptions actually obtained in the census. Unfortunately time is not on your side. Most of the basic material must be collected from earlier applications of the same survey. It is also possible to get basic material from pilot studies and from other surveys where the same variable is under study. However, those latter possibilities might be hazardous.

In fact it is very important that the basic material is up to date. In the Swedish experiments with automatic coding on census material the basic material consisted of schedules from the 1965 censuses. On the basis of that material independent 1970 and 1965 census material concerning industry and occupation have been coded automatically. We found that the coding of the 1965 material was more successful than the coding of the 1970 material. The probable reason for that is a change in the population during these five years. Changes can be structural, i.e. entry and exit of industry and occupation categories occur. It is also possible that the reporting pattern has changed during such a long period of time. One example could be the following: In the 1965 census of population cleaners described their occupation as "cleaner". In the 1970 census a new term, "local keeper", was used by some cleaners. The new term was not even invented in 1965 and as a consequence it was not represented in the basic material. The result was that the dictionary based on the 1965 census material could not code the 1970 census individuals describing their occupation as "local keeper".

Considering the coding error experience shown above in this paper the expert coding of the basic file ought to be verified. For instance, a sequential independent scheme with two experts (and a third when necessary) could be used. The descriptions of the expert coded basic file are of different kinds. We have descriptions with high or low frequencies which point at specific codes. We have variations of these (including abbreviations, spelling errors and so forth) and we have descriptions with high or low frequencies which do not point at specific codes. When we are constructing our dictionary we are interested in covering the first two of these categories. We want to keep the last one out of the dictionary.

The dictionary could be constructed by man or by computer. Presumably a combination of the two is the most efficient approach. In most of our experiments at the Swedish National Central Bureau of Statistics (SCB) the dictionaries have been constructed manually. However, we now have a program working for computerized construction.

The following is a brief description of the manual construction phase.

The expert coded file is first sorted according to code number (list no 1) and after that alfabetically (list no 2). These two lists are the material for the dictionary construction. List no 1 is used to get some hints about the structure of the verbal descriptions sorted under a certain code.

We now choose a frequency limit for classification of "high frequency" descriptions. Then high frequency descriptions are stored in the DA-dictionary (Direct Access), which is scanned first in automatic coding. After that we start looking for discriminating word strings to deal with the variants.

These word strings are stored in a subdictionary called CM (Central Memory). This dictionary is

scanned if the DA-dictionary fails to code a certain description.

By means of list no 2 we check whether the descriptions stored in the dictionaries are unique or not. This check leads to reducing the dictionaries since only unique or "almost unique" descriptions are permitted.

The word strings in the CM-dictionary, which are expensive to look for, should be common to several descriptions or be parts of special highly frequent descriptions.

We have to control that those word strings which are included in the CM-dictionary do not fit the DA-descriptions for other codes. Besides they must be unique in the sense that the same word string does not show up more than once in the CM-dictionary. Unfortunately such controls can not be carried out until a first version of the dictionary is available for each code.

Parts of this job could be carried out by a computer. Such efforts have been shown in O'Reagan (1972) and in Corbett (1972). At the SCB our computerized system contains two programs. One program, LEXSRT, abbreviates the incoming descriptions. After that the descriptions are sorted and the frequency of descriptions with the same code is computed. This file is now used as an input to another program, DALEX, with a couple of sub-routines, CMLEX and CMLIST. DALEX puts the descriptions in the DA-dictionary except for descriptions with low frequency (this value could easily by changed) and for identical descriptions with different codes. In fact we allow "almost unique" cases. We buy coding degree to the price of a hopefully small computer coding error. DALEX calls the sub-routines CMLEX and CMLIST. CMLEX creates an abbreviated description (a six letter word string consisting of the first six letters of the DA-description) and puts it in the CM-dictionary. If the word string is not unique then a new six letter word string is created starting with letter number two in the DA-description. Then the program tries again. At most six such word strings are created. After that the program gives up. CMLIST removes the unusable word strings from the CM-dictionary.

5.2 MATCHING AND CODING

The general matching problem is that exact matchings can be obtained only for a fraction of the verbal descriptions to be coded. We are saved by the fact that for most variables a relatively small number of DA-descriptions is enough to code a relatively large part of the descriptions. For the variants we use the CM-dictionary. Earlier we have used special matching rules. For instance we used a method based on Spearman's rank correlation coefficient. The method worked but the costs were prohibitive.

For automatic coding with the dictionaries described above we use the program AUTKOD. As an input the file with descriptions to be coded is used. Each such description is abbreviated according to the same rules applied when constructing the dictionary. Then the program checks whether the description exists in the DA-dictionary. If so the code is assigned. If not the first six letter word string of the description is matched with the CM-dictionary. If match occurs a code is assigned. If not a new word string is created according to the same rules applied when constructing the CM-dictionary. If match has not occured after six such trials the description is rejected to manual coding.

5.3 SOME EXPERIMENTS AT THE SCB

At the SCB we have carried out automatic coding of the industry variable. The descriptions come from censuses and Labor Force Surveys. This coding has not been especially successful.

Table 5 Automatic coding of industry

Experiment	Kind of dictionary	Kind of data	Coding degree (%)	Quality (% correct coding)
1	Manual	1965 census	50	80
2	Manual	Labor Force 1974	65	69
3	Computerized	1970 census	61	83

Perhaps one can accept the low coding degree but the errors are too frequent. One reason is that the descriptions are rather long for this variable. On the other hand we have not been working with the dictionary that much.

We have been more successful with the occupation variable.

Table 6 Automatic coding of occupation

Experiment	Kind of dictionary	Kind of data	Coding degree (%)	Quality (% correct coding)
1	Manual	1965 census	62	95
2	Manual	1970 census	66	92
3	Manual	1970 census	74	84
4	Manual	1970 census	80	90
5	Manual	Labor Force 1974	81	81
6	Computerized	1970 census	69	87

For census coding we have an acceptable dictionary. The low quality on Labor Force coding is explained by the fact that a translation of the census dictionary was used. Now we have a dictionary based on Labor Force descriptions but it has not yet been tested. The less successful result of the computerized dictionary is explained by the fact that it is still "untouched by human hands". Obviously it is a good raw material for further work.

We have also tried to code goods in the Family Expenditure Survey. The results are good.

Table 7 Automatic coding of goods

Experiment	Kind of dictionary	Kind of data	Coding degree (%)	Quality (% correct coding)
1	Computerized	Family Expenditure Survey 1969	78	93
2	" -	" -	80	93
3	" -	"-	82	96

The results are so promising that automatic coding will be used in the 1978 Family Expenditure Survey.

5.4 GENERAL CONSIDERATIONS

Automatic coding have to be cheaper than manual to be considered. The automatic coding itself is cheap but the punching and the manual coding of the rejects is not. So far we have not been able to calculate costs with enough precision in our experiments. The laboratory differs from reality. However, we are now going to predict the costs for an automatic system in the 1978 Family Expenditure Survey. Manual coding of the whole survey will cost 1,4 million crowns. Automatic coding of the whole survey will cost .07 million. The extra punching of rejected verbal descriptions will cost .2 million. Thus we have quite a margin for manual coding of the 20% rejected and the extra punching of these.

6 REFERENCES

Corbett, J P (1972): Encoding from Free Word Descriptions. U S Bureau of the Census, Draft.

Dalenius, T and Lyberg, L (1968): An Experimental Comparison of Dependent and Independent Verification of Coding, Memo.

Dalenius, T and Frank, O (1968): Control of Classification, Review of the International Statistical Institute, Vol 36:3

Fasteau, H, Ingram, J and Mills, R (1962): Study of the Reliability of Coding of Census Returns, American Statistical Association Proceedings, Social Statistics Section.

Fasteau, H, Ingram, J and Minton, G (1964): Control of Quality of Coding in the 1960 Censuses. Journal of the American Statistical Association, Vol 59, No 305, pp 120-132.

Jabine, T B and Tepping, B J (1973): Controlling the Quality of Occupation and Industry Data, Invited paper to the 1973 ISI meeting.

Lyberg, L (1969): On the Formation of Coding Teams in the Case of Independent Verification under Cost Considerations, Forskningsprojektet "Fel i undersökningar", rapport nr 18, Stockholms universitet.

Lyberg, L (1967): Beroende och oberoende kontroll av kodning, Forskningsprojektet "Fel i undersökningar", rapport nr 4, Stockholms universitet (In Swedish).

Minton, G (1969): Inspection and Correction Error in Data Processing, Journal of The American Statistical Association, pp 1256-1275.

Minton, G (1970): Some Decision Rules for Administrative Applications of Quality Control, Journal of Quality Technology, pp 86-98.

Minton, G (1972): Verification Error in Single Sampling Inspection Plans for Processing Survey Data, Journal of the American Statistical Association, pp 46-54.

Olofsson, P O (1965): PM beträffande variabiliteten i näringsgrens- och yrkesangivelser vid arbetskraftsundersökningar, SCB/UI. (In Swedish).

O'Reagan, R T (1972): Computer-assigned Codes from Verbal Responses. Communications of the ACM, No 6.

U S Bureau of the Census (1972): Coding Performance in the 1970 Census, U S Government Printing Office, Washington, D.C.

FACTORS AFFECTING THE VALIDITY OF SELF-REPORTS OF DRUG USE: AN OVERVIEW

Robert L. Hubbard, William C. Eckerman, J. Valley Rachal and
Jay R. Williams, Research Triangle Institute

Various methods of assessing drug use, abuse and dependence are available, including urinalysis, informants, prescription records, hospital admissions, and arrest reports. None, however, is potentially as useful and accurate as an individual's self-report. Two major obstacles have been identified that may limit the accuracy of self-reports: (1) inability to identify drugs that have been taken, and (2) reluctance to admit a socially undesirable, deviant, or often illegal act.

Two key determinants of a respondent's ability to identify products have been suggested: recognition and recall [24]. Recognition can be defined as knowledge of the name or image of a product that has been used. Recall is the ability to report that the product has been taken at some time in the past. Despite the conceptual distinction, operationally it is difficult to isolate any one factor as the main barrier to accurate identification.

The purpose of this paper is to enumerate and discuss a number of different techniques that have been used to (1) facilitate the recognition of drugs, (2) assist the recall that a particular drug was used, and (3) encourage honest reports of use. Suggestions are presented for methods or combinations of methods to produce the most accurate reports of both past and current use.

RECOGNITION

Two methods of facilitating recognition are: (1) presentation of products in meaningful categories, and (2) use of cues that help a respondent identify products used. Prudent use of both these techniques should increase the validity of self-reports.

Manner of Presentation

Similar products are usually grouped into general categories though few studies use common categories. This is especially true for stimulant and depressant prescription medicines [3,13]. Two distinct ways of presenting products are used. One is based on the pharmacological effect of the product and the other on the way the product can be obtained, by prescription, over-the-counter, or illegally.

There is even a greater problem in categorizing substances that are generally used illegally. For example, in some questionnaires LSD is a separate item; in others, LSD is included under the broader heading of hallucinogens or psychedelics [3]. Most products labeled as hallucinogens such as THC or mescaline actually contain LSD, phencyclidine, or MDA [27]. Thus, the use of separate categories for LSD and other hallucinogens could actually produce an underestimate of the use of LSD.

Despite the advantage of comparability and simplicity in using a series of probes for a few general categories rather than each product used, specificity is sacrificed. In most studies where estimation of patterns of use is only one of many areas of interest, a series of probes for every pill that had ever been used would take far too long.

One study [2] combined both procedures. Respondents were asked which if any pills within a general class of products were used. The followup questions referred to the drug or pill that was used most often or most recently. One of the major aims for future research would be the development of an efficient way to obtain meaningful responses for both individual products and groupings of similar drugs. A more refined procedure might be established where the respondent answers questions about a general class of products and then indicates the specific product or products that he/she had in mind when answering the questions [18].

Aids to Recognition

There are at least six cues that may help a respondent recognize a product: functional descriptors, pharmacological categories, generic names, tradenames, streetnames, and pictures. Some workable combination of these cues can facilitate recognition. Too few aids may not provide enough information to improve recognition. Too many may confuse the respondent, producing higher rates of false positive reports (as when fictitious drugs are listed) or an underreporting.

Functional descriptors indicate reasons for use or effects of use. Functional descriptors should indicate more than the common "upper" or "downer" terms often used [14]. Descriptors such as "to calm down, to relax or to reduce tension" have been used in national studies of psychotropic drug use [28,31]. Functional descriptors may produce more valid reports of general use patterns. However, if a more precise discrimination among products with similar functions (such as sedatives, tranquilizers and barbiturates) is desired, these general cues may confuse respondents. One can first ask a question about a general functional descriptor grouping and then proceed to questions about specific products within that general group. An alternative would be to ask about the use of specific categories of products followed by a question about the use of any other products with similar functions.

Products have been placed in pharmacological categories in any number of ways, often creating confusion. Classification systems with a number of levels have been suggested [8,31]. Data should be collected in a way that permits translation of the results back into generally

accepted pharmacological classification systems. Within a particular category, pictures, functional descriptors or tradenames could be used as cues and examples. The examples should be products that are or were most prevalent in the time period covered by the interview [22,31,34].

Generic names are rarely if ever used in drug use questionnaires or interviews. New guidelines on substitutibility of drugs [39], however, may make generic names more important cues than individual tradenames.

One of the most common cues used by researchers, particularly for prescription products is the tradename of particular pills [2,28,31]. Ninety percent of the pills respondents indicated using in one study were reported by name [24]. The use of tradenames, given the number, variety, and titles [28] does present a number of problems. Physicians and druggists may not tell patients the tradenames of products prescribed or sold [28,30]. Over six pages of products with tradenames so similar they are easily confused by even druggists, nurses and physicians were listed in one report [38].

The problem with the use of street terminology to label illegal substances or products obtained illicitly is even more complex. The Bureau of Narcotics and Dangerous Drugs has listed over 40 terms for marihuana, at least 20 for cocaine and up to 30 for amphetamines. Names may differ across time, regions of the country or within communities in the same metropolitan area. Use of such "vernacular" is often viewed as an "attempt to cozy up to the students" and the terminology for different substances constantly changes [25]. It was concluded that the use of street terms, particularly inappropriate ones, may damage the rapport in the interview.

Another problem with the use of street terminology is that the report of the use of a product does not guarantee that the product was accurately labeled by the distributor [43]. A third of street drug samples analyzed in one study [26], contained substances entirely different from what was advertised. Virtually all street drugs have been found to be falsely labeled or adulterated at some time, including: barbiturates [10], heroin [35], cocaine [36], THC [6,16,27,36] and LSD [27].

The use of visual aids is one technique that has been shown to increase the validity of reports of drug use [30]. Only five percent of the respondents who used drugs in one national sample [28] were not able to identify the name of at least one product they had used with the aid of a color photo chart of products. However for street drugs, different capsules, bootleg chemists, and devious modes of merchandising make meaningful recognition of illegal substances difficult [43]. A forward to a pamphlet showing pictures of the "300 most abused drugs" cautions that "most of the commonly abused drugs are nondescript and therefore extremely hard to identify visually" [4]. The researcher is faced with a problem of how many pictures and which pictures to use as cues. Too many pictures could produce confusion that reduces the validity which might be obtained without pictures. Two examples used in national studies are three cards with approximately eight pictures per card [1] and a chart with over 120 different drugs [31]. Neither method appeared to produce dramatically different results.

RECALL

One study which directly analyzed effects of different factors found that the most important influence on recall was the currency of use [30]. Respondents who had filled prescriptions in the year prior to the interview gave reports of 20 percent greater validity than respondents who last filled prescriptions over a year prior to the interview. In addition, the validity of reports of use of antibiotics in the previous year were almost 20 percent less valid than reports of psychotropic drug use. Since psychotropic drugs are generally used over a longer time and are refilled more often and in greater quantity than antibiotics, it was concluded that the difference in validity may be attributable to the recency and duration of psychotropic drug use.

In a national study of psychotropic drug use [28,31], only five percent of the respondents who used psychotropic drugs were not able to recall the specific name of a product used in the past year. The more recent the use, the more readily the specific product name was recalled.

Respondents may be able to recall using a class of products, but individual products often may be confused even when pictures are provided as cues. Seventeen percent of the respondents who filled prescriptions for stimulants reported instead the use of sedatives or tranquilizers. On the other hand only four percent of respondents filling sedative prescriptions and two percent filling tranquilizer prescriptions reported using products other than these. From the data it is difficult to determine why there was more inaccurate recall by stimulant users.

One hypothesis is that respondents could not recall which pills they had taken. In the national study [28], some respondents who reported using tranquilizers named "aspirin" or "dexedrine" as the specific tranquilizer indicating a problem in recognition. Another hypothesis is that respondents think stimulant, especially amphetamine, use is more deviant or less acceptable than depressant use, indicating a problem in reporting. In another study [15], a number of respondents indicated uncertainty about ever trying a particular kind of product. Not sure responses accounted for almost 10 percent of the answers for five prescription psychotropic products and over 20 percent of the answers for the barbiturate category.

In a comparison study [23], more reports of past and current occasional marihuana use were obtained in self-administered questionnaires than in personal interviews. Fewer reports of frequent past use of marihuana were obtained in

the self-administered questionnaire. It was concluded that the interview procedure may have instigated a more complete recall of past experience.

Another technique that is usually interpreted as a test of the honesty of self-reports, reports of a fictitious drug, may represent a problem in recall. Fictitious products have been included in a number of studies [9,14] that found that very few respondents reported using these products. Followup probes in one study [19] indicated that most respondents reporting the use of fictitious products thought the false drug existed. Rather than indicating a tendency to exaggerate use, two studies [32,42] seem to show that multiple drug users may not be able to accurately recall the types of products used. They may admit the use of a product even if they have some doubt about whether they have taken the product or that the product really exists. The similarity in names of different products [38] may contribute to overreporting, especially among multiple drug users who are exposed to a variety of drugs. Thus, recall may be confounded by the ability to recognize products used.

Specific kinds of products do seem to produce problems in recall. Recent use seems to be the only factor that clearly facilitates recall. In a systematic study of these issues it should be possible to test the effects of different variables by examining the main effects and interactions of these variables on recall of past and current use of different types of products. Covariates of age, education, and sex should also be included in any design. Other factors such as admission of honesty, cooperation in the interview, or ability to comprehend the complexity of the questions should also be considered.

REPORTING

Assuming the product can be identified, a second major concern is whether a respondent will, in fact, report using the product. The possibly threatening nature of the act of reporting use may tend to inhibit completely honest reports [7]. A variety of procedures ranging from simple to complex have been used to elicit reports. However, few attempts have been made to assess the efficacy of one method compared to another. In the following sections some of the procedures typically used will be described and discussed.

Anonymity

Many studies use some procedures to keep responses completely anonymous, protecting not only the respondent but also the researcher [13]. One reviewer [3] reported that a "secret ballot" [29] produced a higher report of use than a personal interview [11] in two national surveys of college students conducted in the same year. However, no substantial differences in reports of use between identifiable and anonymous questionnaires were found in a variety of studies at different colleges [21].

In more controlled studies similar conflicting results were obtained. Clearly anonymous forms did not seem to produce more reports of use than three other types of identifiable forms [14]. In another study [23] the opposite result was found: eight percent more respondents reported using marihuana in the anonymous compared to a coded form.

Overall it appears that no conclusive evidence has been presented that anonymity produces more reports of use. It is possible that if the anonymous nature of the response is overemphasized there may be a "boomerang" effect of increased suspicion. The potential gains of anonymity seem to be outweighed by the advantages of having some way to link drug use reports to other information or to match interviews in succeeding waves of a longitudinal study [14].

Confidentiality

One element that can not be eliminated from an interview is the assurance of the confidentiality of responses. Many procedures have been employed, but no methodological studies have been reported that test the effects of the different methods of assuring confidentiality. In a national survey [1] a self-administered questionnaire was sealed in an envelope by the respondent and could be mailed by the respondent so that interviewers would have no knowledge of responses. In another study [25] materials were sent outside the country where one serial number was removed and a second number was placed on the form. It was felt that these procedures encouraged more cooperation by convincing both interviewers and respondents of the sincerity of the researchers' efforts to maintain confidentiality.

A statewide study of high school students required parental permission for participation and linked respondents to parents and peers only by self-generated code numbers [20]. These procedures resulted in a refusal rate of 14 percent in New York City and less than 50 percent matching of respondents to parents and peers.

One technique that might prove valuable in insuring protection for both the respondent and the researcher is the randomized response technique [40]. This procedure was used with some success in a study of marihuana use and attitudes toward use in a sample of Army enlistees [5]. One problem with this technique is that it is difficult to design probes and formats for followup questions. However, it may be useful for estimating prevalence of illicit drug use.

Interviews versus Self-Administered Questionnaires

The issue of confidentiality raises a critical question of how responses are recorded. At present no clear evidence of greater accuracy of either interviewer administered or self-administered procedures is available. The evidence that is available is unclear or can be interpreted in other ways.

One reviewer [3] hypothesized that despite possible differences in the samples and in response rates, the ten percent difference in reports of marihuana use and the four percent higher report of LSD use may have been due to the self-administered mail questionnaire procedure [29], compared to a personal interview [11]. A problem in selection of different samples for two response conditions confounds the interpretation of the results of another comparative study [23]. Although it was found that reports of frequent past use of marihuana were more prevalent (23 percent) in a personal interview than in anonymous (9 percent) or coded (8 percent) self-administered questionnaires, the samples for each condition differed greatly in size and reason for participation in the study.

Both personal interviews and self-administered questionnaires have been used successfully in a variety of studies. However, in the two National Commission studies [1,2], 10 percent of the respondents in national adult random probability samples could not read the self-administered form and an additional 15 percent appeared to have some trouble reading the form. Based on the results of studies on the effects of assurances of confidentiality [17,21], the number of respondents unwilling to publicly report use may be far smaller than the number who are confused or cannot read the self-administered questionnaire.

Interview Format

The design of the interview schedule could produce motivations to respond more forthrightly to drug questions. How the interview itself is introduced, the context in which the drug questions are embedded, and the order of presentation of the products could influence responses. None of these issues appears to have been empirically evaluated.

Although there are exceptions [41], few if any drug surveys deal only with drug use. Some are introduced as investigations of health [28,31], social issues [1,2,17], or life styles [25]. In validity studies [30,33], if respondents perceive any connection with past history or that records can be checked to verify their responses, they may be more likely to give valid responses.

How the transition to drug questions is made and how it relates to the stated purpose of the study could either increase a respondent's suspicion or reduce a reluctance to respond. For example, one survey [24] introduced drug questions with preliminary questions about a respondent's health problems, symptoms, and means of coping with them. Another survey [17] interspersed drug questions to check the internal consistency of responding with apparently successful results. However, in general, suddenly asking a question about drug use or interspersing drug questions in other contexts in the interview could arouse hostility and consequently lower the validity of use reports.

A third issue in formatting the instrument is the order in which products are presented. Most studies start out with innocuous products such as cigarettes or alcohol, proceed to marihuana and conclude with questions on heroin or opiates. Although intuitively preferable, there is no empirical evidence that this procedure produces more valid responses. Respondents may become more and more defensive as the social undesirability of the products increases. Starting with illegal substances may catch a respondent off guard and initially produce more valid responses, but it may increase defensiveness about answering succeeding questions on the use of objectively less threatening products.

Wording

An often overlooked but critically important aspect of the methodology of constructing a questionnaire on drug use is the wording of the items designed to assess patterns of use. A variety of wordings have been employed for a variety of purposes. Different ways in which items are worded may produce different rates of response.

A very soft wording [11,12] ("Have you ever happened to try ...?") may produce more reports of experimental or one time use. Asking how often a product is used may pick up only current users [25]. In a pretest two respondents admitted "trying" cigarettes; when asked how often they "used" cigarettes, they stated emphatically that they had never "used" cigarettes [15].

A second effect on question wording may be a better estimate of the number of false negative reports of nonuse. More than one category of nonuse, such as the degree of the desire to try the product have been employed in two studies [17,37]. Analyses of such responses could also indicate respondents who may not have reported truthfully. Including a response alternative that permits a respondent to either admit uncertainty about use or evade denying use could indicate the rate of false negative responses, especially those due to problems of recognition or recall. Five percent of the answers in one pretest [15] indicated that respondents were "not sure" they had ever tried particular kinds of products.

CONCLUSIONS

In this paper, we have attempted to present a number of elements to take into account in the assessment of drug use patterns by self-report. Although a variety of approaches and techniques have been suggested, there is little definitive evidence of the impact of any one or any combination of techniques on self-reports of drug use. It is apparent that more systematic methodological studies are needed to identify the most effective ways to obtain self-reports of drug use. These studies need to consider at a minimum the characteristics of the respondent, types of drugs used, and temporal patterns of use as well as the design of the data collection

instrument and the procedures for obtaining the self-reports. An attempt should also be made to compare and integrate the designs and results of the proposed studies with the designs and results of methodological studies of collecting other types of complex and sensitive information through self-reports.

REFERENCES

[1] Abelson, H., Cohen, R., & Schrayer, D. "Public Attitudes Toward Marihuana, Part 1: Main Report." In National Commission on Marihuana and Drug Abuse, *Marihuana: A Signal of Misunderstanding*, Appendix, Vol. II. Washington, D.C.: U.S. Government Printing Office, 1972.

[2] _____, and Rappeport, M. "Drug Experiences, Attitudes and Related Behavior Among Adolescents and Adults." In National Commission on Marihuana and Drug Abuse, *Drug Use in America: Problem in Perspective*, Appendix Vol. I. Washington, D.C.: U.S. Government Printing Office, 1973.

[3] Berg, D.F. "The Non-Medical Use of Dangerous Drugs in the United States: A Comprehensive View." *The International Journal of the Addictions*, 5 (1970), 777-834.

[4] Bludworth, E. *300 Most Abused Drugs. A Pictorial Handbook of Interest to Law Enforcement Officers and Others.* Tampa, Fla.: Trend House, 1969.

[5] Brown, G.H., and Harding, J.C. *A Comparison of Methods of Studying Illicit Drug Usage.* Alexandria, Va.: Human Resources Research Organization, 1973.

[6] Brown, J.K., and Malone, M.H. "Some Street Drug Identification Programs in the United States and Representative Analytic Results." *Journal of the American Pharmaceutical Association*, (1974).

[7] Cannell, C.F., and Fowler, F.J. "A Comparison of Self-Enumerative Procedures and a Personal Interview: A Validity Study." *Public Opinion Quarterly*, 27 (1963), 250-264.

[8] Elinson, J., Haberman, P.W., Hervey, L., and Allyn, A.L. *Operational Definitions of Terms of Drug Use Research.* Report for SAODAP. New York: Columbia University, 1974.

[9] Fejer, D., and Smart, R.G. *Drug Use and Psychological Problems Among Adolescents in a Semi-Rural Area of Ontario: Haldimand County.* Toronto: Addiction Research Foundation, 1971.

[10] Finkle, B.S. "Ubiquitous Reds: A Local Perspective on Secobarbital Abuse." *Clinical Toxicology*, 4 (1971), 253-264.

[11] Gallup Opinion Index. "Results of a Survey of College Students." *Report No. 68.* Princeton, N.J.: Gallup International, February 1971.

[12] _____. "Current Views of College Students on Politics and Drugs." *Report No. 80.* Princeton, N.J.: Gallup International, February 1972.

[13] Glenn, W.A., and Richards, L.G. *Recent Surveys of Nonmedical Drug Use: A Compendium of Abstracts.* Rockville, Md.: National Institute on Drug Abuse, 1974.

[14] Haberman, P.W., Josephson, E., Zanes, A., and Elinson, J. "High School Drug Behavior: A Methodological Report on Pilot Studies." In S. Einstein and S. Allen, *Proceedings of the First International Conference on Student Drug Surveys.* Farmingdale, N.Y.: Baywood Publishing Co., 1972.

[15] Hubbard, R.L. *A Comparison of Required Versus Voluntary Reports of Past and Current Drug Use.* Unpublished Research Report. Milwaukee: The University of Wisconsin, 1975.

[16] Janes, S.H. and Bhatt, S. "Analysis of Street Drugs: A Six Month Study of the Actual Content of Illicit Drug Preparations in a Community." *Journal of Drug Education*, 2 (1972), 197-210.

[17] Johnston, L. *Drugs and American Youth.* Ann Arbor, Mich.: The University of Michigan, Institute for Social Research, 1973.

[18] Johnston, L.D. "Drug Use During and After High School: Results of a National Longitudinal Study." *The American Journal of Public Health* Supplement, Part Two, 64 (1974), 29-37.

[19] Josephson, E. "Trends in Adolescent Marihuana Use." In E. Josephson and E. E. Carroll, *Drug Use: Epidemiological and Sociological Approaches.* Washington, D.C.: Hemisphere Publishing Corp., 1974.

[20] Kandel, D. "Interpersonal Influences on Adolescent Illegal Drug Use." In E. Josephson and E. E. Carroll, *Drug Use: Epidemiological and Sociological Approaches.* Washington, D.C.: Hemisphere Publishing Corp., 1974.

[21] King, F.W. "Anonymous Versus Identifiable Questionnaires in Drug Usage Surveys." *American Psychologist*, 25 (1970), 982-985.

[22] Levy, R., and Brown, A. "An Analysis of Calls to a Drug Crisis Intervention Service." *Journal of Psychedelic Drugs*, 6 (1974), 143-152.

[23] Luetgert, M.J., and Armstrong, A.H. "Methodological Issues in Drug Usage Surveys: Anonymity, Recency and Frequency." *The International Journal of the Addictions*, 8 (1973), 683-689.

[24] Manheimer, D.I., and Mellinger, G.D. "The Psychotropic Pill Taker: Will He Talk?" *Public Opinion Quarterly*, 31 (1967), 436-437.

[25] _____, Somers, R.H., and Kleman, M.T. "Technical and Ethical Considerations in Data Collection." In S. Einstein and S. Allen, *Proceedings of the First International Conference on Student Drug Surveys.* Farmingdale, N.Y.: Baywood Publishing Co., 1972.

[26] Marshman, J.A., and Gibbons, R.J. "The Credibility Gap in the Illicit Drug Market." *Addictions*, 16 (1969), 22-25.

[27] McGlothlin, W.H. "The Epidemiology of Hallucinogenic Drug Use." In E. Josephson and E. E. Carroll, *Drug Use: Epidemiological and Sociological Approaches.* Washington, D.C.: Hemisphere Publishing Corp., 1974.

[28] Mellinger, G.D., Balter, M.D., Parry, H.J., Manheimer, D.I., and Cisin, I.H. "An Overview of Psychotherapeutic Drug Use in the United States." In E. Josephson and E. E. Carroll, *Drug Use: Epidemiological and Sociological Approaches.* Washington, D.C.: Hemisphere Publishing Corp., 1974.

[29] "New Mood on Campus." *Newsweek,* December 29, 1969, 42-45.

[30] Parry, H.J., Balter, M.B., and Cisin, I.H. "Primary Levels of Underreporting Psychotropic Drug Use." *Public Opinion Quarterly,* 34 (1971), 582-592.

[31] _____, Mellinger, G., and Manheimer, D. "National Patterns of Psychotherapeutic Drug Use." *Archives of General Psychiatry,* 28 (1973), 769-783.

[32] Petzel, J.P., Johnson, J.E., and McKillip, J. "Response Bias in Drug Surveys." *Journal of Consulting and Clinical Psychology,* 40 (1973), 437-439.

[33] Robins, L.N., Davis, D.H., and Nurco, D.N. "How Permanent Was Vietnam Drug Addiction." *The American Journal of Public Health,* Supplement, Part Two, 64 (1974), 38-43.

[34] Rucker, T.D. "Drug Use: Data, Sources and Limitations." *Journal of the American Medical Association,* 230 (1974), 888-890.

[35] Schnoll, S.W., Weisman, M., and Lerner, N. "Quality of Street Heroin." *New England Journal of Medicine,* 289 (1973), 698-699.

[36] Smith, D.E. "Street Drug Analysis and Community Based Drug Programs." *Journal of Psychedelic Drugs,* 6 (1974), 153-159.

[37] Tec, N. "Differential Involvement with Marihuana and its Sociocultural Context: A Study of Suburban Youths." *The International Journal of the Addictions,* 7 (1972), 655-669.

[38] Teplitsky, B. "Drug Nomenclature--a Dilemma." *Journal of Drug Issues,* 4 (1974), 135-141.

[39] Tice, L.F. "The Pharmacist and Drug Product Selection." *American Journal of Pharmacy,* 146 (1974), 67-69.

[40] Warner, S.L. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association,* 60 (1965), 63-69.

[41] Whitehead, P.C. *Drug Use Among Adolescent Students in Halifax.* Province of Nova Scotia: Youth Agency, 1969.

[42] _____, and Brook, R. *Social and Drug Using Backgrounds of Drug Users in Treatment: Some Implications for Treatment.* London, Ontario: Addiction Research Foundation, 1971.

[43] _____, and Smart, R.G. "Validity and Reliability of Self-Reported Drug Use." In S. Einstein and S. Allen, *Proceedings of the First International Conference on Student Drug Surveys.* Farmingdale, N.Y.: Baywood Publishing Co., 1972.

THE RANDOM VARIATION IN RATES BASED ON TOTAL ENUMERATION OF EVENTS

J. Richard Udry
Charles Teddlie
C.M. Suchindran
University of North Carolina, Chapel Hill

ABSTRACT

The instability of rates based on total enumeration of events, although not sampling error, may be thought of as being generated by random processes operating in the population. It is therefore necessary to use probability statistics to estimate a "true" rate to determine whether two rates based on total enumeration of events are different from one another. The binomial model has customarily been used to generate predicted variances on the basis of which such determinations are made. Using birth rates for five years from population units of various sizes from Taiwan, North Carolina, and Costa Rica, we obtained empirical estimates of variance in rates which are much larger than those predicted by the binomial model, even when corrections are made for time trends and unit effects. Some of the possible sources of the discrepancy in estimates are discussed. If the binomial model is used to test null hypotheses about the differences in such rates, non-conservative assertions will result.

INTRODUCTION

Social scientists and policy makers are often interested in observing changes in the rate of occurrence of events in discrete population aggregations. For example, one may be interested in knowing whether homicide rates in Pocono County are different for whites and non-whites, or whether the birth rate in a census tract in Manhattan has fallen during the last five years, or whether the motor vehicle accident rates in two counties are significantly different. Such rates are usually derived through complete enumeration of the events rather than sampling, and hence are not subject to sampling variations (errors). Thus, one may think that observed rates pretty much tell the "true" situation. It is well known, however, that the smaller the population base on which such a rate is compiled, the more unstable is the rate over time.

The purpose of this paper is to show how the instability over time of such rates, although not sampling error, may be thought of as being generated by random processes operating in the population. Statistically speaking these events are the outcome of a random experiment. These outcomes (such as birth, death, accident) are subject to chance. Thus, the observed rate may deviate from the "true" rate. Such deviation is called random error. If the experiments are repeated, a measure of this random error can be obtained by obtaining the average deviation of the observed rates around a "true value". Since in this case the experiments cannot be repeated, other procedures have to be developed to obtain measured of random variation. This paper will describe some procedures based on empirical data of birth rates, but the same logic is applicable to the rate of occurrence of many events in a population.

The considerations proposed here are important in many scientific and practical decisions concerning changes in rates. In the development of statistics for small areas, we need to consider the minimum sized population which will provide useful information. In the conduct of field experiments, the investigator often selects small population aggregates as units of "treatment." It is helpful to have a logic for the selection of the size of such units which takes into consideration the random errors in the rates of interest. Any city, considering whether some policy has been effective in changing rates of traffic accidents, crimes, fires, etc., is confronted with the same problems of determining whether the change in rates is "real".

Consider the following hypothetical data from a rural district in Costa Rica containing 5,000 people, and having the following number of births in five consecutive years: 200, 175, 215, 180, 160. For simplicity let us assume that these figures represent the true number of births which occurred. Question: is the birth rate in year five lower than in year one? On the face of it, it seems obvious that the answer is yes. But let us consider that this population contains perhaps 1,000 women of child-bearing age. In any one year about one in five gets pregnant. Which ones? Imagine that the process determing which women get pregnant is stochastic. Some will go through periods of non-exposure to risk, through illness, spouse-absence, etc. Among those exposed during the year, we can imagine pregnancy as a random gift. Whether the birth rate is actually lower in year five than in year one depends not only on the absolute difference between the two rates, but also on the random variation within the rates.

Chiang and Linder[1] seem to have been the first to examine the statistical variations in such vital rates. They have examined random errors and sampling errors of death rates in a variety of situations. They state, "The random error is associated with experimentation, whereas sampling error is due to sampling. These two kinds of error have a subtle but important difference. An understanding of these errors and their difference is essential for the understanding of the standard error of a rate." Keyfitz[2] has discussed the idea of statistical variations of some life table functions. Keyfitz[3] also discussed some measures of random deviations in crude death rates, and direct and indirect adjusted death rates under binomial and poisson

conditions. Walsh[4] and Wilson[5] have also considered some measures of life table death rate and expectation of life at birth. Kupper and Kleinbaum[6] and Kupper[7] have discussed testing equality of k indirect age-adjusted death rates from p($\geq$k) populations in which they derive a measure of random variations of some functions of indirect age specific death rate under the binomial condition. Most of the estimates of measures of random deviation obtained in the above papers make use of binomial or poisson conditions and make several simplifying assumptions to derive them.

An empirical approach to the problem of random variation in vital rates was made by Spencer[8]. She considered the problem of the effect of size of population on variability of demographic data in a historical population. Suchindran et al.[9] approached the problem using Monte Carlo simulation techniques to obtain estimates of random variations in several fertility measures.

These approaches to determining the random variation in demographic rates all assume that probability statistics may be used to estimate the "true" demographic rates. Quite often the binomial model is used to generate these "true" rates. This paper will compare estimates of variability generated by the binomial model with variations found in actual birth rate data to determine how accurate a predictor the binomial model is.

DATA SOURCES AND RESULTS

Selection of Data

The birth rate data come from three separate sources. Annual data for small units were available from North Carolina, Taiwan, and Costa Rica for the 1968-1972 period.[10] These three countries were selected because they had accurate birth rate data for areas as small as 5,000 in population. Seven population size categories were selected: (1) 0-5,000; (2) 5-10,000; (3) 10-15,000; (4) 15-20,000 (5) 20-30,000; (6) 30-40,000; (7) 40-50,000.

Data from fifty-seven cantons in Costa Rica which fell in the 0-50,000 population range in 1968 were used. Originally sixty-two cantons were in this range, but five had to be omitted due to geographic subdivisions during the 1968-1972 period. Data from the four precincts in Taiwan which had at least one township in the 0-5,000 population range were used. This resulted in data from ninety-one townships in Taiwan.

In North Carolina, birth rates were available for whites only, non-whites only, and total combining whites and non-whites. Data for the 0-5,000 population category were based only on non-whites, since white populations exceeded 5,000 in almost all of the counties. Data for the other six categories were based on total rates combining whites and nonwhites. Thus, twenty-two counties using nonwhite birth rates only constituted the 0-5,000 category, while sixty-six counties combining white and nonwhite rates were used in the other six population categories.

Estimates Under the Binomial Model

Table 1 presents a summary of the estimated standard errors based on the binomial model. For each population size category, a mean population (N) was calculated by averaging the 1970 p population of all units within that category. The estimated value of the binomial parameter ($\hat{p}$) for each category was calculated by averaging the crude birth rates over the five year time period of all units within that category. The standard error of the birth rate for each category was then estimated using the formula, $\sigma = \frac{\hat{p}(1-\hat{p})}{N}$, where $\hat{p}$ is the estimated value of the binomial parameter.

As expected, the estimated standard errors decrease as the population size increases. Within each specific population size category, North Carolina shows the smallest estimated error (except for the 0-5,000 category).

Coefficients of variation were calculated by dividing the estimated standard error of each category by the average crude birth rate for that category. These coefficients also decrease as the population size increases. The coefficients of variation for North Carolina are higher than those for Taiwan or Costa Rica, because the crude birth rate is lower in North Carolina than in the other two countries. The estimated coefficients of variation for Taiwan and Costa Rica are very similar.

Observed Standard Errors of Birth Rates

Standard errors based on observed birth rate data from these three countries were calculated next. Assuming time homogeneity, the variance in crude birth rate for each individual unit over the five year time period was calculated (Appendix A, Formula A-1). Next the average variance for each population size category was calculated by averaging the variances of the units within that category under the assumption that rates are unit homogeneous (Appendix A, Formula A-3). Finally, the standard error for each category was calculated by taking the square root of the average variance for that category. These standard errors and their corresponding coefficients of variation are presented in Table 2.

Coefficients of variation based on this analysis generally decrease as the population size increases, but there are a number of exceptions (for example, in Costa Rica at the 20-30,000 level and the 40-50,000 level). The standard errors and coefficients of variation generated from these observed rates are much higher than those that were generated using the binomial distribution.

Observed Standard Errors of Estimate Eliminating Linear Trends

Since there appeared to be a generally decreasing trend for the birth rates over the years observed, a re-analysis was performed. In this analysis, linear trends were eliminated by fitting straight line regressions (Appendix A, Assumption 3). Separate regression lines were fitted for each unit within a population size category, and the mean square error for the deviation from the regression line was calculated

for each unit. The mean square errors of the units within each population size category were then averaged. The standard error of the estimate for each category was then determined by taking the square root of the average mean square error for that category. The results of this analysis are presented in Table 3.

Elimination of linear trends brought about significant reductions in the size of the standard errors of the estimate as compared to the observed standard errors. These reductions are greater in Taiwan and Costa Rica than in North Carolina. Similarly, the coefficients of variation based on the standard errors of the estimate eliminating linear trends are reduced compared to those based on observed standard errors. Despite this reduction, these coefficients of variation and standard errors of the estimate are still consistently higher than those predicted from the binomial model.

Estimates Using Two-way Analysis of Variance

The estimates derived so far have assumed homogeneity of the units within a population category. Under this assumption we have averaged the within unit variation to get a single index of variation. However, when the assumption of homogeneity of units is not satisfied, the true variance of rates will be over-estimated. On the other hand, the process of obtaining separate variances for each unit and then averaging the variances usually results in a reduced estimate of the variance compared to the one obtained by taking a single estimate of variance ignoring the unit classification. These two biases have conflicting effects which may not balance one another.

In order to eliminate both biases, it was decided to re-analyze the data eliminating the assumption of homogeneity of units and the averaging procedure. In this new procedure, a two way analysis of variance was performed with time and units as the two factors. (Appendix A, Assumption 4). This analysis of variance gives an estimate of the random variation in the rates after eliminating the unit and time variations from the total variation. The procedure also allows for testing for trends in time effects, and the deletion of variance associated with linear, quadratic, and cubic time trends.

The analysis was carried out only for Taiwan and the results are presented in Table 4. This analysis revealed that there were significant differences among the units for all the population categories considered. The results also showed that apart from significant linear time trends in all categories, there were four categories with significant quadratic effects and four categories with significant cubic effects. The standard error for a given category presented in Table 4 was obtained by taking the square root of the mean square error for that particular category.

The coefficients of variation in Table 4 show a pattern of decline (with some exceptions) as the population size increases. The estimates in general are larger than the estimates based on the binomial model (Table 1) and smaller than the observed standard errors (Table 2). A comparison of the rates for Taiwan in Tables 3 and 4 shows that the two-way analysis of variance procedure gives smaller estimates in the smaller population size categories and larger estimates in the larger population size categories.

DISCUSSION

The variances generated by actual data consistently display larger values than those predicted by the binomial model for hypothetical populations. We have explored some possible reasons for the over-estimates, but found them unhelpful in reducing the discrepancy. There are other possible reasons which need exploration

1. The binomial model predicts the variance for an infinite number of replications, while we have relatively small numbers of replications available. However, if this were the primary explanation, then the discrepancy between the predicted and the obtained variance should be inversely proportional to the number of replications available. This is not the case, as can be seen from comparison of Tables 1 and 3.

2. We have used crude birth rates, which include in the denominator all persons in the population. However, not everyone in the population is at risk of birth. Denominators should include only the number of women at risk of birth. Birth rates per thousand women at risk would have been preferable, but we did not have these data available. However, this cannot explain the discrepancies. Assuming that perhaps the number of women at risk is one-fifth the number of persons in the population, we would use the reduced denominators in both the empirical estimates and the binomial estimates. The discrepancies would be exactly the same size, but the size of population to which they applied would be one-fifth as large.

3. The mean square error of the crude birth rate (which is the square of the standard error of the estimate) is equal to the true error plus the correlation between error and time. If there is a correlation between error and time (a circumstance which we can rarely evaluate), the standard error of the estimate would be slightly larger than the true error.

4. The simple binomial model assumes that every woman is at the same risk of birth. Surely this is an erroneous assumption. If one assumes that the risk of birth varies, then the simple binomial model will underestimate the variance in birth rates. Consider the following example.

Suppose a population of 1,000 women with the probability (p) of a birth in a year of .01. Assuming p is constant for all women, the expected number of births in a year is 10, and the variance is equal to 1,000 x .01x.99=9.9, or variance in the birth rate of 9.9/1,000. Now assume that p varies among women with a mean of .01, and a variance of only .00001. The expected number of births is still 10, but the variance is now 19.9/1,000. (See Appendix B for the equation). By adding a very small variance to p, we have more than doubled the variance in the birth rate.

If we could decompose any population into sub-populations with the same probability of experiencing the criterion event, our estimates

would probably more closely approach those predicted by the binomial model. But from a practical point of view, this observation is of little assistance, since the circumstances under which we can either decompose the population into groups with the same p, or alternately, estimate the variance of p, are extremely unusual. Assuming that we are usually dealing with populations in which p has some unknown distribution, our predicted variances based on the simple binomial model seem doomed to be over-estimates.

SUMMARY AND CONCLUSIONS

This paper has explored the estimation of random variation in rates based on total enumeration of events. It is not concerned with variations due to sampling and response errors. Assessment of random variation in rates is necessary to detect changes with time as well as differentials in rates between regions or groups. It is necessary to determine minimum sample size needed to detect change or differentials, or minimum change in rates which cannot be attributed to random factors. It is necessary in establishing the size of statistical reporting units which will provide sufficiently stable rates for various purposes.

Several measures of random variation are presented. The variance generated by the most widely used binomial model displayed smaller values than any of those generated by our empirical data. We have identified difficult-to-eliminate sources of random variance which may make any empirically derived variance estimates substantially larger than those predicted by the binomial model. The use of the binomial model to estimate predicted variances against which to test null hypotheses can therefore be expected routinely to result in the rejection of null hypotheses which should in fact have been accepted. It will therefore lead to nonconservative assertions of true differences in rates where none in fact exist. If the experience with birth rates in other populations and the experience with other types of rates is similar to that we have presented, conservative inferences will require estimates of predicted variances made from detailed data on the actual population being studied.

APPENDIX A
Measures of Random Variation

Let b_{tk} denote the rate at time t (t = 1,2, . . .s) and for unit k (k = 1,2, . . .ℓ).

The following measures of random variation can be obtained.

Assumption 1. Rates are time homogeneous

A measure of random variation for unit k is given by (A. 1) $S^2_{1k} = \frac{1}{s-1} \Sigma(b_{tk} - \bar{b}_k)^2$, when $\bar{b}_k = \frac{1}{s} \Sigma b_{tk}$

Assumption 2. Rates are time and unit homogeneous

The following measures of random variations can be constructed.

(A.2) $S^2_2 = \frac{1}{\ell s-1} \Sigma_t \Sigma_k (b_{tk} - \bar{b})^2$ when $\bar{b} = \frac{1}{tk} \Sigma\Sigma b_{tk}$

(A.3) $S^2_3 = \frac{1}{\ell} \Sigma S_{1k})^2$

(A.4) $S^2_4 = \frac{(1}{\ell} \Sigma S_{1k})^2$

Assumption 3. Assume that the rates change with time. $b_{t,k} = B_{o,k} + B_{1,k}t + B_{2K}t^2 + \ldots + B_{2k}th + E_k$, when E_k is $N(0, \sigma^2)$.

Then an estimate of the variance of observed $b_{t,k}$ is given by the mean square error for the deviation from the best fitted regression line.

If all units are assumed to be homogeneous, then an improved estimate can be obtained by taking an average of the standard error obtained for each region.

Assumption 4. Rates are not homogeneous with respect to time and region. In this case, it is better to eliminate region and time effects from the total variation of the rates. This can be done using the analysis of variance technique. Using orthogonal polynomials one can also test for the linear, quadratic, cubin etc.. time trends of the rates. (For a standard reference, see Snedecor and Cochran.[11])

An estimate of the variance of the rate is obtained from the mean squares due to error in the analysis of variance table.

APPENDIX B
Variance in Binomial Model

Assume that p is the probability of occurrence of an event in a year for a member of the population. Then, for a population of size N, the observed rate will have an expected value of p, and variance p(1-p)/N.

Now assume that p varies among women with mean value of p* and variance σ^2_p. Then, it can be shown that the observed rate has an expected value of p* and variance equal to

$$V_{\hat{p}} = \frac{1}{N^2} [Np^*(1 - p^*) + N(N - 1)\sigma^2_p]$$

Note that, when N is large, the second term of the sum does not disappear.

FOOTNOTES

*Partial support for this project was provided through grants from the Ford Foundation and the National Institute of Child Health and Human Development, (Grant #HD05798) to the Carolina Population Center.

1. C.L. Chiang and F.E. Linder, "On the Standard Errors of Death Rates", (mimeographed) Population Laboratories, University of North Carolina at Chapel Hill, 1969.

2. N. Keyfitz, "Sampling Variance of Demographic Characteristics", Human Biology, 38, 1966, pp. 22-41.

3. N. Keyfitz, Introduction to the Mathematics of Population, Addison-Wesley, 1968.

4. J.E. Walsh, "Large Sample Tests and Confidence Intervals for Mortality Rates", Journal of the American Statistical Association, 45, 1950, pp. 225-237.

5. E.B. Wilson, "The Standard Deviation of Sampling for Life Expectancy", Journal of the American Statistical Association, 33, 1938, pp. 705-708.

6. L.L. Kupper and D.G. Kleinbaum, "On Testing Hypothesis Concerning Standardized Mortality Ratios", Theoretical Population Biology, 2, 1971, pp. 290-298.

7. L.L. Kupper, "Some Further Remarks on Testing Hypothesis Concerning Standardized Mortality Ratios", Theoretical Population Biology, 2, pp. 431-436.

8. B. Spencer, "Size of Population and Variability, of Demographic Data (17th - 18th centuries)." Paper presented at the annual meetings of the Population Association of America, April, 1975.

9. C.M. Suchindran, J.W. Lingmer, A.N. Sirha, and E.J. Clark, "Sensitivity of Alternative Fertility Indices," Proceedings of the Social Statistics Section of the American Statistical Association 1976, pp. 798-805, Washington, D.C.

10. The birth rate data from North Carolina were obtained from North Carolina Vital Statistics 1968, 1969, 1970, 1971, and 1972 published by the North Carolina State Board of Health, Public Health Statistics Division. Data from Costa Rica were obtained from the Republica de Costa Rica Estadistica Vital 1968, 1969, 1970, 1971, and 1972 published by the Departamento Estadisticas Sociales, Seccion Estadistica Vital. The Taiwanese data were obtained from Taiwan Demographic Fact Book 1968, 1969, 1970, 1971, and 1972 published by the Ministry of the Interior of the Republic of China.

11. G.W. Snedecor and W.G. Cochran, Statistical Methods, The Iowa State University, 1968.

TABLE 1. Estimated standard errors of birth rates based on binomial model

Data source	Population size category	Number of units	Average crude birth rate	Estimated standard error	Coefficient of variation
Costa Rica	0 - 5,000	4	23.6	2.216	9.4%
	5 - 10,000	12	29.5	1.903	6.5%
	10 - 15,000	22	32.4	1.526	4.7%
	15 - 20,000	7	34.4	1.360	4.0%
	20 - 30,000	6	32.0	1.050	3.3%
	30 - 40,000	2	33.6	.942	2.8%
	40 - 50,000	4	34.6	.820	2.4%
North Carolina	0 - 5,000	22	22.9	2.894	12.6%
	5 - 10,000	11	16.2	1.487	9.2%
	10 - 15,000	10	16.4	1.101	6.7%
	15 - 20,000	12	17.2	.946	5.5%
	20 - 30,000	17	17.9	.818	4.6%
	30 - 40,000	7	18.0	.703	3.9%
	40 - 50,000	9	17.7	.610	3.4%
Taiwan	0 - 5,000	14	33.2	3.010	9.1%
	5 - 10,000	10	29.7	1.934	6.5%
	10 - 15,000	12	29.7	1.455	4.9%
	15 - 20,000	19	29.3	1.263	4.3%
	20 - 30,000	20	28.0	1.045	3.7%
	30 - 40,000	9	28.9	.856	3.0%
	40 - 50,000	7	29.2	.785	2.7%

TABLE 2. Observed standard errors of birth rates

Data source	Population size category	Number of units	Average crude birth rate	Average standard error	Coefficient of variation
Costa Rica	0 - 5,000	4	23.6	4.375	18.5%
	5 - 10,000	12	29.5	4.782	16.2%
	10 - 15,000	22	32.4	3.593	11.1%
	15 - 20,000	7	34.4	3.577	10.4%
	20 - 30,000	6	32.0	3.485	10.9%
	30 - 40,000	2	33.6	2.910	8.7%
	40 - 50,000	4	34.6	5.551	16.0%
North Carolina	0 - 5,000	22	22.9	4.799	21.0%
	5 - 10,000	11	16.2	1.807	11.2%
	10 - 15,000	10	16.4	1.832	11.2%
	15 - 20,000	12	17.2	1.963	11.4%
	20 - 30,000	17	17.9	1.131	6.4%
	30 - 40,000	7	18.0	1.133	6.3%
	40 - 50,000	9	17.7	0.976	5.5%
Taiwan	0 - 5,000	14	33.2	4.082	12.3%
	5 - 10,000	10	29.7	3.171	10.7%
	10 - 15,000	12	29.7	2.338	7.9%
	15 - 20,000	19	29.3	2.951	10.1%
	20 - 30,000	20	28.0	2.649	9.5%
	30 - 40,000	9	28.9	2.217	7.7%
	40 - 50,000	7	29.2	2.345	8.0%

TABLE 3. Observed standard errors of estimate of birth rates eliminating linear trends

Data source	Population size category	Number of units	Average crude birth rate	Average standard error of the estimate	Coefficient of variation
Costa Rica	0 - 5,000	4	23.6	2.496	10.6%
	5 - 10,000	12	29.5	2.438	8.3%
	10 - 15,000	22	32.4	2.344	7.2%
	15 - 20,000	7	34.4	2.397	7.0%
	20 - 30,000	6	32.0	1.611	5.0%
	30 - 40,000	2	33.6	1.322	3.9%
	40 - 50,000	4	34.6	1.758	5.1%
North Carolina	0 - 5,000	22	22.9	3.910	17.1%
	5 - 10,000	11	16.2	1.564	9.7%
	10 - 15,000	10	16.4	1.444	8.8%
	15 - 20,000	12	17.2	1.552	9.0%
	20 - 30,000	17	17.9	0.987	5.5%
	30 - 40,000	7	18.0	1.062	5.9%
	40 - 50,000	9	17.7	0.806	4.6%
Taiwan	0 - 5,000	14	33.2	3.667	11.0%
	5 - 10,000	10	29.7	2.167	7.3%
	10 - 15,000	12	29.7	1.463	4.9%
	15 - 20,000	19	29.3	1.731	5.9%
	20 - 30,000	20	28.0	1.258	4.5%
	30 - 40,000	9	28.9	1.230	4.3%
	40 - 50,000	7	29.2	1.140	3.9%

TABLE 4. Observed standard errors of estimate for Taiwan birth rates eliminating time and unit effects

Population size category	Number of units	Average crude birth rate	Standard error	Coefficient of variation
0 - 5,000	14	33.2	3.558	10.7%
5 - 10,000	10	29.7	2.147	7.2%
10 - 15,000	12	29.7	1.390	4.7%
15 - 20,000	19	29.3	1.718	5.9%
20 - 30,000	20	28.0	1.533	5.5%
30 - 40,000	9	28.9	1.290	4.5%
40 - 50,000	7	29.2	1.365	4.7%

LOG-LINEAR ANALYSIS OF FERTILITY USING CENSUS AND SURVEY DATA WITH AN EXAMPLE

I. Elaine Allen and Roger C. Avery, Cornell University

The use of log-linear models is relatively new to the field of demography, especially in the analysis of fertility. Previous log-linear analyses have been largely studies of cohort mobility and of infant mortality [8]. This paper shows the value in fitting multiway tables when analyzing fertility with census and survey data. An example of differential current fertility using a 20% sample of the census of Costa Rica will be presented. The variables were all discrete and categorical which made the log-linear approach an appropriate technique. In addition, with this method we had the potential of creating as detailed a contingency table as was necessary and could evaluate the complicated interaction terms in a simple, systematic, and statistically robust manner.

Often research in fertility presents findings in the form of tables. Two and three way tables have often been used but the construction of higher than three-way tables becomes difficult to synthesize and unwieldy to present in tabular form. This type of analysis misses a great deal concerning the factors influencing fertility. Even if chi-square statistics are calculated for the respective tables it is usually difficult to identify a unifying strain within many multiway tables. Also, without a systematic method of constructing tables to include patterns of interactions between variables, there can be little comparability between them.

One method of getting around this has been linear regression. Historically fertility analysis has dealt with aggregate measures or small samples and was thus appropriate for regression analysis. Individual level analysis for large samples has generally not been possible as data has been unreliable or incomplete on the individual level. Adding interaction terms to regression models is possible using dummy variables but, unlike log-linear models, there is no simple way to identify and test the interaction terms in the model. While any combination of variables can be input in a log-linear model with one term, many terms are required for interactions in dummy variable regression.

Again, returning to aggregate level analysis, another often used method is the construction of various fertility rates for comparison within a crosstabulation by the variable of interest. While this gives good comparisons within variables it becomes a cumbersome procedure as the categories of the variables and the dimensions of the table increase. It provides no overall measure of the significance within and between these rates, so conclusions based on these rates alone may be tenuous. Also, the rates for a country may differ greatly from those of smaller areas within the country or with the individual.

There are several applications of log linear models, not all are of interest here. For example, one may wish to examine the fitted table as well as the Likelihood Ratio Statistic. However, in this paper we are more interested in how well the model fits the table, reflected in the goodness of fit statistics and in the fitted parameters of the model. Rather than simply finding the table's best fit we are interested in a dependent variable approach. In our approach we are interested in identifying the factors that effect the distribution of the dependent variables within each cell, and the direction and strength of those factors, rather than factors that effect the number of cases in each cell. It has been shown by Goodman [9] that when analyzing a contingency table using a dependent variable approach only those terms involving the dependent variable need be included as all other terms will cancel.

Therefore, we are concerned with sampling from a Multinomial distribution where the population being studied can fall into one and only one of t categories with a probability p_i where (p_i) is the vector of cell probabilities summing to one for the t categories [5]. The p_i reflect the relative frequency of each category in the population. A structure may be imposed when using two or more variables or dimensions, the data are usually represented as groups of rectangular arrays. This structure can be described by models linear in the logarithmic scale. The term model used here is analogous to the equation of linear regression; its parameters, additive and multiplicative effects, are similar to metric beta coefficients and their significance, and the significance of the whole model is measured by the magnitude, or goodness of fit, of the Likelihood Ratio Statistic. The lack of fit of the model may be compared to the magnitude of the error sum of squares in regression or inversely to the multiple-R^2.

We are interested in the amount of reduction in the Likelihood Ratio Statistic occurring between two models, which gives an indication of the importance of adding an additional term to the model. The statistic reported in fitting a multiway table gives an indication of the fit of the entire model to the observed data while the difference between statistics indicates the importance of individual terms. The models fit in this paper are hierarchical models. High order terms may only be included in the model if the related low order terms are included [9]. In assessing the significance of any particular term or interaction to the model several measures of

association are available. The two measures used for testing particular terms in this paper were marginal and partial association [10]. These show the effect of adding a higher order term to the saturated model of next lowest order and the effect of dropping a term from the model of a certain order, respectively.

Two 10% samples of the 1973 census of Costa Rica were available, these were non-overlapping systematic samples of families and were combined for the purposes of this work. From this sample, a file was created for each woman over 15. Using methods similar to those developed by Lee-Jay Cho and others of the East-West Center the number and ages of own children were estimated for these women. Own children are children present in the family who cannot be shown not to be a woman's children [3]. The relative ages of the woman and child and the number of surviving children a woman had were used as criteria in this process. If she had more children present than surviving the oldest children were assumed not to be own children [3]. Women 50 and older who were not likely to have had children in the five years before the census, the widowed and divorced, and women under 20 were excluded from the sample, yielding 87,540 cases.

Two models were fit, the first used the number of children born to each woman in the five years before the census as its dependent variable. From other analysis it can be shown that this variable gives a good approximation of period fertility rates in Costa Rica on an overall basis. A second model used a dichotomized version of this variable, dividing women into those who did or did not have children in the five year period. The use of two models allowed a thorough examination of the information that was lost in this dichotomization.

The independent variables were: Age, in five year age groups; Marital Status, Single, Married or Consensual Union; Urban /Rural; Education, None, Primary and Secondary or more; Working/Not Working. Since the complete fertility histories were available through the use of the own children method we could develop a control variable based on the woman's fertiltiy history at the beginning of the five year period in question. Five categories of previous fertility were delineated ranging from those with no previous children to those with eight or more children. The inclusion of this variable as a control had two purposes: It allows us to distinguish timing patterns as we are, in effect, explaining the change in fertility from one period to the next and it also gives an indication of the variables not included in the model by how important a part it plays in determining differential fertility.

Results

Tables 1, 2, and 3 give the detailed results of fitting the models. The L.R.S. of fitting both models is given in Table 1. Though both fit well the model with six categories of fertility was slightly better. A small, and not significant L. R.S. indicates a small difference between the observed and expected values and is desireable for a well fitting model. Both models include only those term of the third order with age or previous fertility controlled. For example the interaction of Urban/Rural, Age and Current Fertility and the interaction of Urban/Rural, Previous Fertility and Current Fertility were included. Each term was tested and marginal and partial association were both significant. The possibly confounding effect of the interactions between independent variables has been controlled for by including the six-way interaction of all the independent variables.

Tables 2 and 3 present the multiplicative parameters of the model, first with dichotomous current fertility and next with the full six categories. In Table 2 the second order effects appear across the top of the table and in the right-most column. These are the interactions of Current Fertility and each independent variable. The body of the table contains the third-order effects, or the indirect effects; interactions of Current Fertility and each independent variable controlled for Age or Previous Fertility. In Table 3 the second order effects are in 3a and the third order effects controlling for Age and Previous Fertility are in 3b and 3c respectively. The multiplicative parameters for the dichotomous dependent variable are reciprocals of one another and for both models these parameters are constrained to multiply to 1 within any category. It is the parameters' difference from 1 which determines how great an effect it is having on the dependent variable. In examining Table 2, for example, the second order effect for Urban/Rural is .788 on experiencing current fertility and 1.268 for no fertility. This variable has a fairly strong effect on Current Fertility here but when the effect of Age is controlled, in 2, the third order Urban/Rural multiplicative effects are close to 1.

From these tables we see that fertility in rural Costa Rica, and for women in Consensual Unions, is higher than that of urban areas or women who are married, especially in younger age groups. The more education a woman has, or if she is working, the fewer children she has. The differential is greatest at younger ages and reverses at the higher ages perhaps showing that these women have merely postponed childbearing while working or going to school or perhaps because of their higher social status.

Although the results of fitting the models are complex, a significant pattern emerges: for social and economic variables the differentials decrease with age. This

is the opposite of what would be expected with the demographic transition which supposes that the differentials in fertility depend on the age at which women cease childbearing. The patterns of three way interactions for Previous Fertility are, in a sense, reversed from those by age, the zero parity women have the smallest differentials by Education and Urban/ Rural while the high parity women have strong interactions. This is as expected by the theory of demographic transition. Childlessness is a function of exogenous factors such as sterility while, the parity at which women stop childbearing is expected to be affected by her social class.

The log-linear model with six categories of current fertility has the effect of taking the women who had experienced fertility in the last five years and further dividing them by fertility. While we found distinct advantages to the dichotomized variable, among them the ease of presentation of one number for each category of each independent variable and the ability to easily calculate the total odds ratios of experiencing current fertility from tables and , information about the details of the distribution of current fertility is lost using this variable. Particularly the curvilinear effect of some of the variables on current fertility was not evident when the variable was dichotomized.

In Table 3 patterns can be seen by looking down the categories of current fertility. The overall effect of Urban women experiencing fertility was negative in Table 2 but it is positive for current fertility of one child and highly negative thereafter. So it is large numbers of women having small families and perhaps a family planning mechanism at work. Also, in Table 2, the large positive effects for current fertility 5+ controlled for age and previous fertility reflect a small number of cases in the whole sample. So while the odds of a woman in these categories experiencing high current fertility are great, there are very few women in these categories.

Another relationship exhibited in the six category model is the curvilinear one. Both the single women, those women who are working, and the highly educated are most likely to have no children or a great many children as can be seen in Table 3. For single women, those with many children may be from low status groups or widowed or separated from consensual unions. In the case of the highly educated or working, these are probabily high status women.

The construction of graphs of the multiplicative parameters can be useful for polytomous dependent variables. These can show the spread of the differentials and how they change when controlled by a third variable.

When analyzing fertility in this manner an estimate of a woman's average fertility can be constructed using the odds ratios. After construction of the odds ratio for a certain set of independent variable categories these categories can be converted into probabilities by the relationship

$$P = \frac{Odds}{1 + Odds}$$

and standardized for the number and value of the categories. After multiplying the probability by the current fertility category it is summed and averaged to give the average fertility. For example for Age (20-24); Urban; Single; Not Working; Education (Primary); with no Previous Fertility the probabilities for each category of Current Fertility are as follows:

0 : .8321
1 : .1364
2 : .0305
3 : .0028
4 : .0002
5+: .00003

So the average fertility experienced by a woman in this category for the last five years would be: 0(.8321) + 1(.1364) + 2(.0305) + 3(.0028) + 4(.0002) + 5.3(.00003) = .200.

Conclusions

As we have shown, there are a variety of techniques for presentation of the results of fitting a log-linear model which are meaningful to the demographer: Tables of second and third order effects reflecting linear and curvilinear trends, graphs illustrating the comparison of the differential effects of different variables on fertility before and after controlling for age and previous fertility. the construction of odds ratios and manipulating them to find the probability of being in a certain category of current fertility and finally, taking the product of these probabilities by their fertility category and averaging them to find a measure of average fertility for women in a certain group of independent variable categories.

Log-linear analysis seems especially appropriate for census and survey data for several reasons: The size of the data set can by quite large and the use of regression techniques, especially in evaluating the significance of coefficients, is difficult; The content of the variables is often categorical; The construction and evaluation of interaction is simple and straightforward in log-linear analysis; The alternative to fitting the model, the examination of high order contingency tables, give no significance of interactions or overall fit.

Since so many demographic analyses begin, and sometimes end, with the construction of multiway tables the implementation of a log-linear model is an easy and appropriate step forward from the present methods. Its greatest advantages are in allowing for the determination

and inclusion of interactions of independent and dependent variables and in summarizing what might otherwise by a contingency table of unmanageable size.

References

[1] Allen, I.E. and R. Avery. 1977. "Trends and Differentials in Fertility in Costa Rica using an Own Children Method," Paper presented at the Annual Meeting of the Population Association of America, St.Louis, Mo.

[2] Avery, Roger.1976a."Estimation of Individual Fertility Histories Using Own Children Present, Children Ever Born and Children Surviving, With Examples from Costa Rica," Second Own Children Method Workshop, East-West Population Center, Honolulu, Hawaii.

[3] Avery, Roger.1976b."The Use of the Age Distribution of Own Children in Estimated Childhood Survival," Second Own Children Method Workshop, East-West Population Center, Honolulu, Hawaii.

[4] Avery, Roger.1977."A Comparison of Birth Rates Estimated from the Vital Stat-System of Costa Rica with Birth Rates Estimated from Own Children Methods," Unpublished manuscript, International Population Program, Cornell University, July 1977.

[5] Bishop, Y.M.N. et al. 1975. Discrete Multivariate Analysis. MIT Press, Cambridge, Mass.

[6] Brown, Mort. 1976. "Screening Effects for Multiway Tables," Applied Statistics.

[7] Census Nacional de Costa Rica, 1974.

[8] Goodman, L.A. "On the Statistical Analysis of Mobility." Amer.Jnl. Sociology 1965, 70:564-585.

[9] Goodman, L.A. 1973."Causal Analysis of Data from Panel Studies and Other Kinds of Surveys," Amer. Jnl. Soc. 78:1135-1191.

[10] Stycos, J.M. 1977. "Recent Trends in Latin American Fertility," eds. W. Peterson and L.H. Day. Harvard University Press, Cambridge, Mass.

Acknowledgements

This work was supported in part by NSF Contract dcr75-13373, awarded to the Department of Economic and Social Statistics, Cornell University; and a grant from A.I.D. to the International Population Program at Cornell University, G-1493. The two 10% samples of the 1973 census of Costa Rica from the Latin American Data Bank, Gainesville, Florida and the United States Census Bureau, respectively, with the kind permission of the Direccion General de Estadistica y Censos, San Jose, Costa Rica. The authors wish to thank Jeff Seaman for helpful discussions and comments on earlier drafts.

Table 1: Likelihood Ratio Statistics for Second and Third Order Models Controlled For Age and Previous Fertility.

Dichotomous Current Fertility:

	L.R.S.	d.f.
All second order	8274.38**	1693
Controlled for Age	6740.57**	1245
Controlled for Prev.Fer.	5632.47**	1545
Controlled for both	1016.04*	821

Six Categories of Current Fertility:

All Second Order	12880.20**	5233
Controlled for Age	8590.63**	4637
Controlled for Prev.Fer.	7603.91**	4637
Controlled for Both	2646.97	4105

** .001 level of significance
* .1 level of significance

Table 2. Dichotomous Current Fertility* Controlled for Age

Age of Woman	20-24	25-29	30-34	35-39	40-44	First Order Current Fert. Tau	Prev. Fert. 0	1	2-4	5-7	8+	First Order Curr. Fert. Tau^2
Second Order Tau^2	1.915	1.646	1.367	.780	.498	.978	.490	1.057	1.281	1.203	1.257	.978

Table 2: Dichotomous Current Fertility* Controlled for Previous Fertility

Third Order Tau^2 Interactions Curr. Fert./Age						Second Order Tau^2	Third Order Tau^2 Interactions Current Fertility/Prev. Fertility					Second Order Tau^2
Urban/Rural	U	.953	1.000	1.036	.994	1.022	.788	Urban/Rural	U	1.371	1.052	.882	.897	.876	.788
	R	1.050	1.000	.965	1.006	.978	1.268		R	.729	.950	1.134	1.115	1.141	1.268
Marital Status	S	.885	.939	.982	1.073	1.138	.457	Marital Status	Single	.340	.869	1.423	1.662	1.435	.457
	M	.819	1.036	1.138	1.026	1.008	1.428		Married	2.193	1.270	.738	.635	.764	1.428
	C.U.	1.376	1.028	.895	.908	.872	1.533		Cons. Union	1.341	.906	.953	.946	.912	1.533
Labor Force Status	Non. W.	1.162	1.192	.976	.885	.837	1.237	Labor Force Status	Not W.	.859	1.012	.984	1.089	1.057	1.237
	W.	.861	.819	1.024	1.129	1.194	.809		Working	1.164	.988	1.016	.902	.946	.809
Education	None	1.230	1.1025	.945	.982	.796	.972	Education	None	.837	.846	1.086	1.158	1.119	.972
	Prim.	1.293	1.012	.924	.878	.945	.958		Primary	1.049	.941	.933	1.105	.984	.958
	Sec.+	.629	.897	1.147	1.160	1.332	1.073		Secondary+	1.138	1.254	.988	.781	.908	1.073
Prev. Fert.	0	1.278	1.381	1.002	.711	.774	.490	Age	20-24	1.277	1.646	1.638	.679	.468	1.915
	1	1.646	1.376	.939	.792	.594	1.057		25-29	1.381	1.376	1.096	.974	.493	1.646
	2-4	1.634	1.096	.805	.880	.789	1.281		30-34	1.002	.939	.803	1.008	1.313	1.367
	5-7	.679	.974	1.008	1.234	1.214	1.203		35-39	.733	.794	.880	1.234	1.654	.780
	8+	.428	.493	1.313	1.593	2.271	1.257		40-44	.774	.594	.789	1.214	2.25	.298

*the numbers in the table reflect the odds of experiencing current fertility, the odds of experiencing no fertility are the reciprocals of these numbers.

Table 3 : Current Fertility with Six Categories

Current Fertility		0	1	2	3	4	5+
First Order Tau		12.138	4.752	1.737	.476	.190	.110
Second Order Tau Effects							
Urban/Rural	Urban	1.700	1.320	.891	.677	.796	.927
	Rural	.588	.757	1.122	1.476	1.257	1.079
Marital Status	Single	2.846	.848	.554	.529	.943	1.501
	Married	.910	1.698	1.682	1.369	.723	.389
	Cons. Union	.386	.694	1.073	1.383	1.467	1.711
Labor Force Status	Not Working	.901	1.141	1.570	1.325	.834	.563
	Working	1.110	.877	.637	.755	1.200	1.780
Education	None	.458	.601	.876	1.281	1.766	1.833
	Primary	1.503	1.600	1.362	1.223	.663	.377
	Secondary+	1.450	1.040	.839	.638	.854	1.450
Prev. Fert.	0	4.951	1.173	.947	.701	.537	.482
	1	1.090	1.171	1.008	1.057	.867	.850
	2-4	1.318	1.682	1.311	.826	.587	.709
	5-7	.536	.790	.880	1.084	1.605	1.543
	8+	.262	.548	.908	1.508	2.277	2.232
Age	20-24	.213	.738	1.304	1.748	1.750	1.595
	25-29	.482	1.103	1.418	1.316	1.016	.992
	30-34	.893	1.295	1.462	.841	.863	.811
	35-39	1.902	1.195	.857	.814	.815	.773
	40-44	5.717	.801	.432	.634	.799	1.008

Table 3a: Six Categories of Current Fertility

Third Order Tau Effects Controlled for Age						
Age Group		20-24	25-29	30-34	35-39	40-44
Current Fertility						
Urban:	0	1.032	.970	.939	1.057	1.010
	1	.990	1.026	1.036	1.026	.927
	2	1.069	1.062	1.018	.933	.927
	3	1.004	1.145	.943	.939	.984
	4	.994	.978	.962	.990	1.080
	5+	.918	.846	1.113	1 064	1.085
Marital Status:						
Single:	0	1.621	1.280	1.054	.752	.588
	1	1.243	1.156	1.239	.808	.696
	2	1.355	1.117	.848	1.022	.762
	3	1.201	.843	.731	.908	1.489
	4	.557	.895	.955	1.281	1.644
	5+	.549	.805	1.248	1.383	1.309
Married:	0	.889	.743	.824	1.219	1.510
	1	.679	.803	1.004	1.237	1.484
	2	.630	.863	1.234	1.237	1.206
	3	.884	1.221	1.237	1.032	.728
	4	1.823	1.171	.922	.753	.676
	5+	1.636	1.364	.859	.691	.755
Consensual Union:	0	.696	1.055	1.111	1.092	1.126
	1	1.186	1.080	.805	1.002	.968
	2	1.173	1.038	.955	.792	1.088
	3	.728	.974	1.105	1.067	.925
	4	.986	.955	1.136	1.036	.901
	5+	1.113	.910	.933	1.047	1.012
Labor Force Status:**	0	.699	.740	.978	1.219	1.623
	1	.845	.956	.914	1.049	1.295
	2	1.113	1.100	.904	.929	.972
	3	1.156	1.186	1.145	.914	.697
	4	1.223	1.259	1.012	.874	.736
	5+	1.080	.861	1.069	1.055	.955
Education: None:	0	.687	.790	1.057	1.069	1.631
	1	.759	.887	.945	1.160	1.355
	2	1.272	.785	.889	1.059	1.063
	3	1.107	1.201	.976	.966	.797
	4	1.435	1.055	.988	.988	.764
	5+	1.075	1.435	1.164	.797	.699

*Tau effects for Rural are the reciprocals of those for Urban

** Tau effects for Working are the reciprocals of those for Not Working.

Table 3a (continued)

Age Group		20-24	25-29	30-34	35-39	40-44
Primary:	0	.529	.837	1.130	1.364	1.469
	1	.861	.841	1.026	1.042	1.290
	2	1.042	.972	.893	1.006	1.100
	3	1.362	.876	1.033	1.115	.728
	4	1.266	1.452	1.067	.748	.682
	5+	1.223	1.149	.876	.839	.968
Secondary+:	0	2.752	1.512	.837	.686	.419
	1	1.532	1.341	1.030	.826	.572
	2	.755	1.311	1.259	.939	.856
	3	.664	.949	.992	.927	1.724
	4	.623	.653	.949	1.354	1.918
	5+	.760	.607	.980	1.496	1.473
Previous Fertility	0	3.602	1.111	.805	.706	.440
	1	4.718	1.823	.903	.407	.317
	2	2.443	1.669	.834	.587	.500
	3	.714	1.203	1.115	.970	1.077
	4	.386	.517	.920	1.785	3.049
	5+	.088	.475	1.613	3.426	4.364
1:	0	.960	1.010	1.171	.904	.916
	1	1.362	1.742	1.430	.750	.394
	2	1.538	1.651	1.098	.697	.514
	3	1.358	1.173	.734	.850	.984
	4	.699	.697	.712	1.348	2.140
	5+	.514	.421	1.038	1.737	2.560
2-4:	0	.475	.856	1.237	1.374	1.450
	1	.707	.889	1.042	1.416	1.077
	2	1.062	1.092	1.022	1.169	.724
	3	1.049	1.421	.935	.764	.880
	4	1.223	1.024	1.057	.880	.859
	5+	2.187	.826	.771	.656	1.096
5-7:	0	.486	.990	1.069	1.190	1.633
	1	.374	.796	.870	1.685	2.289
	2	.531	.659	1.126	1.550	1.636
	3	1.243	.933	1.016	1.212	.701
	4	2.100	1.166	1.245	.701	.468
	5+	3.956	1.772	.755	.379	.500
8+:	0	1.254	1.051	.805	.901	1.049
	1	.587	.445	.856	1.376	3.251
	2	.471	.506	.949	1.350	3.283
	3	.776	.534	1.286	1.311	1.433
	4	1.445	2.323	1.162	.674	.381
	5+	2.576	3.415	1.026	.676	.164

*Tau effects for Rural are the reciprocals of those for Urban

** Tau effects for Working are the reciprocals of those for Not Working.

Table 3b Six Categories of Current Fertility

Third Order Tau Effects Controlled for Previous Fertility

Previous Fertility Group:		0	1	2-4	5-7	8+
Current Fertility						
Urban:*	0	.686	1.000	1.932	1.084	.968
	1	1.105	1.077	1.130	.972	.766
	2	1.230	1.128	1.028	.755	.927
	3	1.177	.962	.953	1.032	.867
	4	1.002	.921	.834	1.080	1.203
	5+	.910	.929	.778	1.130	1.348
Marital Status:						
Single:	0	8.952	1.484	.551	.361	.379
	1	1.197	1.194	.958	.893	.817
	2	.640	.845	1.286	1.325	1.086
	3	.494	.908	1.259	1.385	1.279
	4	.508	.951	1.096	1.208	1.563
	5+	.582	.774	1.071	1.395	1.486
Married:	0	.274	.769	1.667	2.019	1.409
	1	1.169	1.212	1.042	.861	.785
	2	1.623	1.189	.771	.741	.908
	3	1.583	1.047	.760	.878	.904
	4	1.177	.815	1.067	.859	1.138
	5+	1.034	1.059	.922	1.028	.966
Consensual Union:	0	.048	.876	1.089	1.371	1.871
	1	.716	.691	1.002	1.300	1.558
	2	.964	.998	1.010	1.018	1.014
	3	1.281	1.053	1.046	.821	.863
	4	1.671	1.293	.854	.964	.563
	5+	1.662	1.221	1.014	.697	.697
Labor Force Status:**	0	1.115	.956	.992	.920	1.028
	1	.887	1.071	.996	1.024	1.032
	2	.897	.939	.968	1.098	1.117
	3	.943	.933	1.030	1.049	1.053
	4	1.179	1.069	1.077	.958	.769
	5+	1.014	1.042	.941	.964	1.042
Education:						
None	0	.879	.964	.760	1.049	1.479
	1	.751	.908	.750	1.203	1.623
	2	.632	.774	1.184	1.279	1.346
	3	.899	.897	1.026	1.128	1.073
	4	1.360	1.210	1.186	.835	.613
	5+	1.957	1.357	1.217	.658	.417
Primary:	0	.931	1.164	1.014	.908	1.004
	1	1.053	.878	.925	1.184	.990
	2	.954	.984	1.006	1.075	.984
	3	1.087	1.006	1.151	1.036	.766
	4	1.121	1.094	1.012	.796	1.012
	5+	.878	.904	.910	1.051	1.318
Secondary+:	0	1.221	.891	1.297	1.051	.674
	1	1.263	1.254	1.440	.702	.623
	2	1.656	1.311	.839	.728	.755
	3	1.024	1.109	.846	.857	1.217
	4	.656	.755	.834	1.503	1.613
	5+	.582	.815	.904	1.447	1.613

* Tau effects for Rural are the reciprocals of those for Urban.

**Tau effects for Working are the reciprocals of those for Not Working.

A COMPARISON OF MALE OCCUPATION SPECIFIC LABOR FORCE SEPARATIONS OBTAINED THROUGH A LONGITUDINAL STUDY AND THOSE OBTAINED THROUGH STANDARD WORKING LIFE TABLES

Ali Rashid, Central Department of Statistics
Kingdom of Saudi Arabia
and
Michael P. McElroy, United States Bureau of the Census/
U. S. Representation to the United States -
Saudi Arabian Joint Commission on Economic Cooperation

Labor Force Replacement Needs

The future occupational employment needs of a labor market area can be forecast by considering expansion needs caused by growth in the total number of jobs in the economy and by evaluating replacement needs for the people holding jobs during the period of study. The vast majority of time and effort by manpower analysts has been devoted to estimating the expansion needs of labor markets. Studies have indicated, however, that the number of job openings resulting from replacement needs frequently exceeds the number resulting from expansion needs. The number of job openings from labor force separations alone in the United States until 1985 is expected to be double the growth openings. 1/

Replacement needs are caused by people leaving the labor force, transferring occupations, and transferring to other localities. This paper will focus on the methods used to estimate needs created when males leave the labor force. These needs are called labor force separations or sometimes deaths and retirements, since they are the usual means of exiting from the labor force.

Method of Computing Male Occupational Labor Force Separations

There are several ways to approach the estimation of male occupational labor force separations. The method applied to a given labor market depends to a great extent upon the type of data available to the analyst. If one knows only the overall percentage of males leaving the labor force each year, then the total number of job openings for a future year can be estimated. This method would have the limitation of assuming the same separation rate for all males regardless of what occupation they were, as well as the assumption that the percentage for the base year will continue in the future. The problem in finding a method of computing male separations has been one of attempting to minimize the number of quantitatively significant limitations.

The most widely used method of estimating male labor force separations is by the use of working life tables and occupation specific age distributions. This method has been used for some time by the U. S. Bureau of Labor Statistics in estimating National and State labor force separations. Age-specific labor force participation rates and age-specific death rates are used to create the working life table.

The tables of working life follow through successive ages the labor force participation experience of the population. 2/ A separate estimate is yielded of deaths and retirements for each age group under consideration. The death rate is the overall death rate for the population obtained from standard life tables. There are several advantages to this method. The main one is that it integrates data sources effectively into one system. A complete explanation of working life tables may be found in BLS Bulletin 1001. 3/

One assumption of the working life table technique is that before the maximum age of participation, no one leaves the labor force except through death. This assumes that there are no disability retirements and also assumes that no one withdraws from the labor force to attend school fulltime. How severe a limitation this is depends on the socio-economic factors affecting the area being considered.

The principal limitation of the working life table/ occupational age distribution method is the assumption that within specific age groups the separation rate does not vary by occupation. It is well known that retirement patterns and even mortality rates do vary by occupation. This assumption is usually detailed when presenting data from the working life table method. There is no way for analysts to circumvent this limitation without tapping another data source.

One additional method of estimating the number of occupational labor force separation openings is by the use of a longitudinal study. The actual labor force participation patterns of a sample can be followed over a time period to determine such factors as deaths and retirements by occupation, occupational transfers, and geographic mobility. This method has the advantage of producing a great deal of information without

having to make adjustments for comparability with other sources.

The longitudinal method while yielding valuable information, however, has problems of data collection practicality. The cost of this type of survey can be great if the sample size is large. Moreover, obtaining respondent cooperation over an extended time can be difficult, particularly since similar information must be asked several times.

Purpose of this Paper

This paper will show the differences that occurred between computing occupational labor force separations by using a longitudinal study and by computing separations from age distributions and a working life table. The source of all data used in the comparison is the Saudi Arabian Labor Force Survey conducted by the Central Department of Statistics. No independent sources were interjected into the comparison that would create a need for adjustment factors to achieve the data comparability. The results will provide analysts with a quantitative measure of the limitations of the standard working life table approach.

Saudi Arabian Labor Force Survey

In order to present a clear explanation of the procedures used in developing occupational separations from the labor force using both methods, it will be helpful to give a brief explanation of the survey from which the data are taken. The Saudi Arabian Labor Force Survey is part of the Kingdom's Multipurpose Household Survey. 4/ Households are contacted for a 13-month period in order to collect information on demographic characteristics, labor force status, and income level. Approximately 60,000 persons were in the survey. Statistics on sampling errors will be available after the final round of the survey is completed.

Major labor force questions are asked two times during the survey period of all household members 12 years of age and over. The survey produces information on employment status, hours worked, occupation, industry, class of worker, job-related income, occupation worked at last year, time since last worked for those not now employed, activity engaged in to find work for those seeking work, last job of the labor force reserve, and second job of dual jobholders. In addition to these major questions, monthly questions are asked to monitor seasonal fluctuations in economic activity. Any labor force information on the individual can be readily classified by age, sex, nationality, educational level and marital status. Information from the demographic characteristics, such as migrations and deaths, are linked by the computer to the labor force information for the individual.

How the Labor Force Separations were Developed - Longitudinal Method

The first step in developing occupation specific labor force separations was to separate all of the labor force participants according to occupational groups. The four groups chosen were: (1) Professional, Technical and Managerial workers; (2) Sales, Service and Clerical workers; (3) Operatives, Laborers and Production workers; and (4) Farmers and workers not classified by occupation. Each of these four groups were further divided into five groups: Under 25 years of age; 25-34 years of age; 35-44 years of age; 45-54 years of age; and 55 years of age and over. Thus, twenty distinct occupational age groups were stratified.

The occupation specific male labor force separations from the longitudinal method were derived by dividing the number of males who were not in the labor force at the end of the survey by the number who were in the survey at the beginning. As of the writing of this paper, nine months of data have been processed, so adjustments were made to produce an annual rate. An exit rate for each of the four occupation groups was developed to compare with the rate obtained through the working life table approach.

Ancillary products of the Saudi Arabian longitudinal study are the development of occupational mobility and geographic mobility estimates. These factors were considered in the separation study since they affect the size of the cohort groups under consideration. If a person changed occupations during the survey period, the occupation he was engaged in during the initial survey round was the one assigned to him for the separation study. Anyone who in-migrated to or out-migrated from a survey household during the survey period was excluded from the labor force separation estimates since their labor force status while out of the survey is unknown. The geographic mobility factor is particularly important to Saudi Arabia due to the large number of foreign workers in the Kingdom.

How the Working Life Table Rates were Developed

A working life table was developed from the data collected through the Multipurpose Survey. The standard methodology found in the BLS Bulletin 1001 was followed in developing the table. Some columns of the standard working life table, such as the average number of remaining years of work, were not developed, as this information

would not contribute to the development of the age-specific labor force withdrawal rates.

The age-specific withdrawal rates produced by the table were adjusted to conform to the five age groups of the longitudinal study. The withdrawal rates by age were then applied to their respective occupational age numbers to arrive at a number of withdrawals for each of the four occupation groups, by the five age groups. A withdrawal rate for each of the four occupational groups was obtained by adding the withdrawals of the five age groups within the occupation and dividing this sum by the total number of males in the occupation.

Comparison of Results

Table 1 (immediately after references) depicts the separation rates for the four occupation groups resulting from both methods.

The Working life table/age distribution labor force withdrawal rate for Professional, Technical, and Managerial people was 182% of the withdrawal rate produced through the longitudinal study. Utilization of the Working life table/age distribution rate would result in a tremendous overestimation of the labor market needs for this occupational group. It was expected that the withdrawal rate for this group would be less using the longitudinal study. The magnitude of the difference, however, was alarming.

The Working life table/age distribution labor force withdrawal rate for Production workers, Operatives, and Laborers was 60% of the longitudinal study rate. The fact that the nature of the work of people in this group leads to more labor force withdrawals within age groups than other occupation groups is not being considered by the working life table approach. The severity of the differences between the Working life table/age distribution approach and the longitudinal studies was not as great in the other two occupational groups. The Working life table/age distribution method produced a rate of 121% of the longitudinal study for Clerks, Salespersons, and Service workers. The withdrawal rate for Farmers obtained through the Working life table approach was 93% of the longitudinal study rate.

The rates obtained in the Saudi Arabian study are unique to that Kingdom, but the numerical discrepancies with the longitudinal study reveal the inadequacies of the Working life table/age distribution approach. Certainly more effort should be directed toward improving the quality of data used to estimate labor force withdrawals.

The number of occupational withdrawals, together with the number of expansion openings, give the labor market analyst an estimate of the total job openings by occupation in an area. Vocational education coordinators frequently use this information to formulate training programs. If a vocational coordinator in Saudi Arabia utilized the results of the working life method to estimate how many people would have to be trained to replace the Production workers, Operatives, and Laborers who exited from the labor force, they would have accounted for only 60% of the labor market needs for these replacement workers.

A more critical situation would be encountered if the coordinator used the working life estimates for Professional, Technical, and Managerial workers. Numerous people would have been channeled into lengthy training programs in anticipation of jobs that would not have existed. Thus, it is essential that analysts continue to seek more effective methods to utilize in estimating labor force withdrawals.

This paper does not serve to render obsolete all applications of the Working life table approach. It has the desirable feature of being able to produce projected rates for future years which is important in analyzing future occupational needs.

As mentioned previously, the longitudinal method has potential data collection problems. The problem of respondent irritance during extended surveys can be minimized, however, if the group conducting the survey takes measures to inform respondents of the purposes of the survey. The officials at the Saudi Arabian Central Department of Statistics spent a great deal of time and effort publicizing the surveys, and these efforts have paid off in a virtually non-existent non-response problem.

References

1) U. S. Bureau of Labor Statistics, Tomorrow's Manpower Needs, Supplement No. 4, "Estimating Occupational Separations from the Labor Force for States." (Washington: Government Printing Office, 1974.)

2) U. S. Bureau of Labor Statistics, Tomorrow's Manpower Needs, Volume One, "Developing Area Manpower Projections, Bulletin 1609." (Washington: Government Printing Office, 1969.)

3) U. S. Bureau of Labor Statistics, Bulletin 1001. (Washington: Government Printing Office, 1950.)

4) Rashid, A. and Rumford, J. "The Multi-purpose Survey of Saudi Arabia - An Experiment in Compression." Proceedings of the American Statistical Association, 1976.

TABLE 1.

Occupation Group	Working Life Separation Rate	Longitudinal Separation Rate
Professionals, Technical, and Managerial workers	.0186	.0102
Clerical, Sales, and Service workers	.0251	.0207
Production workers, Operatives, and Laborers	.0137	.0229
Farmers and workers not elsewhere classified	.0539	.0579

A MODEL OF POPULATION GROWTH INVOLVING MORTALITY FERTILITY INTERACTIONS: PROJECTIONS FOR INDIA

G. K. Kripalani and Rodney Smith, Western Michigan University

A population growth model investigating the implications of exogenous continuously improving mortality experiences in low-income world for sequential fertility changes as lagged response to mortality disturbances and for future population growth and structure was presented at the 1976 Annual Meetings. [See 1976 Proceedings of the Social Statistics Section, pages 501-06]. Some tentative results for India were also presented.

More extensive computer simulation results for several values of the lag parameter are now presented. Additionally, lag parameter values based on relevant population census and other demographic statistics for India have been tentatively estimated and population projections made for the years 1991 and 2011. The future course of important population structural characteristics like dependency ratio, proportion in labor force age-groups, proportion of children in the population, proportion of females in childbearing age-groups, total fertility rate, and long-term stable population growth rate up to year 2011 has been calculated.

It may be worthwhile to restate the basic elements of the analysis underlying this study. Mortality changes are assumed exogenous. The initial population is regarded as a stable population at time t=o. This population becomes subject to exogenously determined rates of mortality improvement of varying magnitude over the next several periods. The central hypothesis is that birth rates may respond in downward fashion to declines in death rates. The main elements of the hypothesis pertain to household family formation behaviour and are: (a) the concept of Desired Family Size; (b) household response to past mortality changes via lagged adjustment in planned fertility; (c) 'myopic' expectations about future mortality improvements; (d) possible changes in (i) desired family size, (ii) preferred child-spacing pattern and (iii) household behaviour parameters reflecting degree of risk-aversion in response to mortality improvements and the historical consistency of this process. The expectations hypothesis involves distributed lags and myopic expectations. Mathematically the hypothesis used is written as:

$$E\,y\,(t + c/t) = M\,(t + c/t).\ E\,y\,(t/t) \quad \ldots.(1)$$

where

$$E\,y\,(t/t) = L\ E\,y\,(t - 1/t - 1) + (1 - L)\,y\,(t - 1) \quad \ldots.(2)$$

where L = lag parameter lying between 0 and 1;

$E\,y\,(t + c/t)$ = expected change in the force of mortality in the period $(t + a)$, expectations formed at time t; and

$y\,(t - 1)$ = actual change in the force of mortality observed in the time period $(t - 1)$.

$M\,(t + c/t)$ = Myopia factor at time t for time period $(t + c)$ in the future.

Changes in mortality rates play an important role in this model on account of the concept of the Desired Completed Family Size and its fixity in the face of changes in mortality. Declining mortality rates and the taking into account of mortality improvements in the decision-making process for determining planned fertility rates imply that planned fertility rates respond to changes in mortality via number of currently living children and expected survival rates. Decline in mortality rates will induce declines in planned fertility rates in order to achieve the goal of a fixed DCFS.

A simplified formulation is developed for the purpose of gaining qualitative insights into the role played by model parameters and for throwing into sharp focus the relationship between fertility and mortality rates in determining age composition structure and rate of population growth. The population is divided into four equal age groups 0, 1, 2 and 3. Age group 0 relates to children and age group 1 consists of all adults in childbearing period of life. Children are born to females in age group 1 only. Since all children are born in one time period, myopia is absent. The myopia parameter $M\,(t + c/t + c)$ is equal to unity.

Family formation behaviour assumptions are: (i) The family is aiming at a Desired Completed Family Size (DCFS) which is assumed given and fixed and does not change as mortality rates change. DCFS is defined as the number of children born who are desired to survive to adulthood, say age 1. (ii) The family has a fixed preferred child-spacing pattern which does not change as mortality and fertility changes occur. (iii) Families respond to mortality improvements by lagged adjustments in planned fertility. Since a single period covers the whole child-bearing time span, it will be unrealistic to ignore completely mortality changes currently under way

whose impact on emerging profile of children living at various ages of the mother's child-bearing span could easily be visible.

Empirical Results for India

For reasons of space, a detailed discussion of the choice of parameter values and of the assumptions underlying the projections is not given here. The following information based on results of 1951, 1961 and 1971 Population Censuses of India is, however, important in making judgments about these assumed values.

(a) The percent growth rates of India's population during 1941-50, 1951-60 and 1961-70 decades were 13.4%, 21.64% and 24.57%. Between 1951 and 1971, India's population increased by 51.1 percent.

(b) If it is assumed that no significant mortality improvements occurred in India in the few decades prior to 1951 so that stable population condition could be taken as a reasonably rough approximation, the long-run stable population one period (20 years) growth factor G may be assumed at $(1.134)^2 = 1.286$. This means that on average, in the absence of significant mortality improvements that actually occurred in India during the fifties and to a much lesser extent during the sixties, India's population between 1951 and 1971 would have increased by 28.6%. The difference of 22.5% may be attributed to mortality and fertility shifts that may have taken place during the 20-year period 1951-71.

(c) Analyses of India's census data suggests that there is little evidence of significant fertility declines occurring during 1951-70 in response to very significant mortality declines underway in that period. This means that the value of lag parameter L in relation (3) is very close to unity.

(d) Based on India's Official Life Tables, the survival rates from birth to age 20 are as follows:

Period	Male	Female
1941-50	.58	.57
1951-60	.72	.71
1961-70	.77	.75

Thus, between 1946 and 1956 (mid-points of the decades), the female's 20-year survival rate increased by 24.56 percent; the percentage for period between 1956 and 1966 was only 5.92 percent. For the 20-year period 1946 to 1966, the 20-year female survival rate increased by 31.93 percent. Evidence is very clear that mortality declines which were very significant during the fifties had considerably slowed down during the sixties. Mortality gains reflected in the above survival rate were of the order of 2.2 percent per year in fifties, but only of 0.6 percent per year in the sixties.

(e) Life expectancy at birth for females was 35 years based on 1941-50 Life Table, 40.0 on 1951-60 Life Table, and 45.6 years on 1961-70 Life Table. Thus, over the 20-years between 1951 and 1971 Censuses, female life expectancy at birth increased by over 10 years, or by nearly 30 percent.

The following assumptions have been made in making population projections:

(i) Calculations have been made for females only. It is assumed that similar orders of magnitude will emerge for males and total population. 50% of children born are assumed female.

(ii) Mortality disturbance is assumed to start at time t=o in 1951. Three different sets of assumptions regarding future mortality improvements over the 3 time periods are used in making projections: a(t) means that forces of mortality at all ages decline on average during time period t to t+1 by amount a(t).

(a) $a(o) = .2820$; $a(1) = 0$; $a(2) = 0$ [once-for-all disturbance case]. [Low Mortality gains case]

(b) $a(o) = .30$; $a(1) = .10$; $a(2) = .05$; [Intermediate mortality gains case]

(c) $a(o) = .32$; $a(1) = .16$; $a(2) = .08$; [High mortality gains case]

Future mortality gains are assumed to be smaller since existing cheap sources of mortality declines are assumed to have been, by and large, almost entirely used up, and further gains are likely to depend on improvements in diet, nutrition, etc.; that is, factors which depend on gains in per capita income.

(iii) For making population projections Model Life Tables West-Females for Life Expectancy at Birth equal 35 years given in Coale and Demeny have been used. [Page 38; r = .10; Mortality Level 7.]

(iv) Tentative analysis of census data and other relevant population statistics for India for the period 1950 to 1970 indicate that the value of the lag parameter relevant for India is close to unity, may be around .90. But there is some evidence that significant fertility reductions as a result of a vigorous Government policy for population control may be underway. Hence, we may use the values of lag parameter as .75 and .9 for population projection purposes. Besides projections are also made for the case of no fertility response L = 1.0 and a value of L = .6 to show the population growth and structure implications of more intensified population control efforts to accelerate fertility response to the lag parameter value of L = .6.

Simulation Results

The main results for the Intermediate

Summary Table

Projections for Female Population for India for Years 1991 and 2010
and Estimation of Important Population Parameters.
Intermediate Mortality Gains Case, a(o) = .30; a(1) = .10; a(2) = .05
(Initial 1971 census figure assumed at 1000)

	Actual 1971	Projected Population							
		1991				2011			
		L=.6	L=.75	L=.9	L=1.0	L=.6	L=.75	L=.9	L=1.0
A. Female Population by age-group									
0 (0-20)	506	543	557	575	589	664	685	718	748
1 (20-40)	286	507	514	518	521	679	696	716	731
2 (40-60)	148	320	314	311	309	572	577	583	584
3 (60-80)	60	122	121	120	118	235	230	226	224
4 Total	1000	1492	1506	1524	1537	2150	2188	2243	2287
B. Proportion age group 0.	.506	.364	.370	.377	.383	.309	.313	.320	.327
C. Proportion age group 1.	.286	.340	.341	.340	.339	.316	.318	.319	.320
D. Proportion "labor force" (=age groups 1 and 2)	.434	.554	.550	.544	.540	.582	.582	.579	.575
E. Dependency Ratio. [(0)+(3)] / [(1)+(2))]	1.304	.804	.818	.837	.853	.717	.720	.728	.739
F. (i) Estimated total fertility (female children only)	2.53 (1951)	1.771	1.780	1.845	1.870	1.624	1.634	1.664	1.695
(ii) Fertility as proportion of 1951 fertility	1	.700	.711	.727	.741	.642	.646	.658	.670
G. Projected Population given actual 1971 total population (millions)	458	683	690	698	704	985	1002	1027	1047
H. Rate of Population Increase (%)									
(i) over period	-	49.2	50.6	52.4	53.7	44.1	45.3	47.2	48.8
(ii) annualized rate	-	2.02	2.07	2.13	2.17	1.84	1.89	1.95	2.01
I. Projected Population for year 2001	-	-	-	-	-	820	832	847	859

Future Mortality Gains case (Case (b)) are given in the Table below. The results based on High and Low cases (cases (a) and (c)) are given in the Table below. The discussion below is based on Case (b).

Important results are:

(i) If the hypothesis that households' fertility behavior takes no account of mortality gains during the current period is true, i.e., L = 1.0, then India's expected population is expected to be 704 million by 1991, 859 million by 2001, and 1047 million by 2011.

(ii) India's population increased by 51.6 percent during 1951-71; it is projected to grow by 49.2 percent if L = .6, 50.6 percent if L = .75, and 52.4 percent if L = .9. The projected rates of growth for the period 1991-2011 are 44.1% (L = .6), 45.3% (L = .75), and 47.2% (L = .90).

(iii) India's child population age group (0-20) which formed 50.6% in 1971 is projected to fall to 36.4% in 1991 and 30.9% in 2011 if L = .6; to 37.7 in 1991, and 32.0% in 2011 if L = .9.

(iv) Total fertility, i.e., number of children born per potential mother, is expected to fall to 70.0% of its 1951 level by 1991 and to 64.2% of its 1951 level by 2011 if L = .6. These work out to 30 percent decline in fertility by 1991 and 36 percent decline by 2011. For L = .9, the fertility declines by 1991 and 2011 are projected to be 27.3 percent and 34.2 percent below 1951 levels.

(v) The proportion of the female population in the child bearing ages is expected to increase from 28.6 percent in 1961, to 34.0 percent in 1991, and to 31.6 percent in 2011 if L = .6. This proportion remains fairly stable for different values of L being around 34 percent for 1991, and 32 percent for 2011.

(vi) The proportion of population in labor force age groups 1 and 2 is expected to rise from 43.4 percent in 1971, to between 44 to 45 percent in 1991, and to 58 percent in 2011. This proportion shows minor variations for different values of L.

(vii) The dependency ratio is projected to decline from 1.3 in 1971 to between .80 and .84 in 1991, and to about .72 in 2001.

References & Bibliography

Coale, A. J.: The Growth and Structure of Human Populations, Princeton University Press, Princeton, N. J., 1972.

Coale, A. J. and Demeny, P.: Regional Model Life Tables and Stable Populations, Princeton University Press, Princeton, 1966.

Hoover, P. F. and Longley-Cook, L. H.: Life and Other Contingencies, University Press, Cambridge, England, 1953. Published for the Institute of Actuaries and the Faculty of Actuaries.

Milbank Memorial Fund: "The Interrelations of Demographic, Economic and Social Problems in Selected Underdeveloped Areas." New York, 1954. Registrar General of India: Official Population Census Reports, 1951, 1961 and 1971.

Registrar General of India: Official Life Tables for India, 1941-50, 1951-60 and 1961-70.

World Bank: Population Planning-Sector Working Paper, World Bank, March 1972.

Acknowledgments

This research is a part of larger research of the first author and Dr. G. S. Tolley of the Univeristy of Chicago on the Economic-Demographic interactions in India's Agricultural Development. Mr. Rodney Smith has been responsible for the final computer simulation work. Earlier work in the writing of computer program and of initial calculations was done by Mr. Alec E. Gores. Both Mr. Smith and Mr. Gores received financial support for participating in this research from Russell H. Seibert Fund of the Honors College of the Western Michigan University.

Annexe A
Table 2
Projections for Female Population for India for Years 1991 and 2011
and Estimation of Important Population Parameters.
High Mortality Gains Case- a(0) = .32; a(1) = .16; a(2) = .08
(Initial 1971 census figure assumed at 1000)

	Actual 1971	Projected Population							
		1991				2011			
		L=.6	L=.75	L=.9	L=1.0	L=.6	L=.75	L=.9	L=1.0
A. Female Population by age-group									
0 (0-20)	506	546	563	584	601	661	687	729	766
1 (20-40)	286	529	535	540	544	743	766	792	813
2 (40-60)	148	333	329	325	322	650	654	659	662
3 (60-80)	60	129	127	126	124	267	263	258	254
4 Total	1000	1537	1554	1575	1591	2321	2370	2438	2495
B. Proportion age group 0.	.506	.355	.362	.371	.378	.285	.290	.299	.307
C. Proportion age group 1.	.286	.344	.344	.343	.342	.320	.323	.325	.326
D. Proportion "labor force" (=age groups 1 and 2)	.434	.561	.556	.549	.544	.600	.599	.595	.591
E. Dependency Ratio. [(0)+(3)] / [(1)+(2)]	1.304	.781	.798	.820	.838	.666	.669	.680	.692
F. (i) Estimated total fertility (female children only)	2.53 (1951)	1.713	1.748	1.796	1.837	1.478	1.493	1.528	1.566
(ii) Fertility as proportion of 1951 fertility	1	.677	.691	.710	.726	.584	.590	.604	.619
G. Projected Population given actual 1971 total population (millions)	458	704	712	721	729	1063	1085	1117	1143
H. Rate of Population Increase (%)									
(i) over period	-	53.7	55.4	57.5	59.0	51.0	52.5	54.8	56.8
(ii) annualized rate	-	2.17	2.23	2.30	2.35	2.08	2.13	2.21	2.27
I. Projected Population for year 2001	-	-	-	-	-	865	879	897	913

Annexe A
Table 1

Projections for Female Population for India for Years 1991 and 2010
and Estimation of Important Population Parameters.
Low Mortality Gains Case - (once-for-all mortality disturbance). a(o) = .2820; a(1) = 0; a(2) = 0.
(Initial 1971 census figure assumed at 1000)

	Actual 1971	Projected Population							
		1991				2011			
		L=.6	L=.75	L=.9	L=1.0	L=.6	L=.75	L=.9	L=1.0
A. Female Population by age-group									
0 (0-20)	506	534	547	557	566	665	681	696	715
1 (20-40)	286	481	489	490	493	590	603	611	622
2 (40-60)	148	300	298	293	291	478	486	488	487
3 (60-80)	60	116	116	113	113	195	193	188	189
4 Total	1000	1431	1450	1453	1463	1928	1963	1983	2013
B. Proportion children [age group 0] .	.506	.373	.377	.383	.387	.345	.347	.351	.355
C. Proportion women child-bearing age [age group 1] .	.286	.336	.337	.337	.337	.306	.307	.308	.309
D. Proportion "labor force" (=age groups 1 and 2)	.434	.546	.543	.539	.536	.554	.555	.554	.551
E. Dependency Ratio. [(0)+(3)] / [(1)+(2)]	1.304	.832	.840	.854	.865	.804	.802	.806	.813
F. (i) Estimated total fertility (female children only)	2.53 (1951)	1.844	1.860	1.885	1.908	1.870	1.870	1.887	1.908
(ii) Fertility as proportion of 1951 fertility	1	.729	.735	.745	.754	.739	.739	.746	.754
G. Projected Population given actual 1971 total population (millions)	458	655	664	665	670	883	899	908	922
H. Rate of Population Increase (%)									
(i) over period	-	43.1	45.0	45.3	46.3	34.7	35.4	36.5	37.6
(ii) annualized rate	-	1.81	1.88	1.89	1.92	1.50	1.53	1.57	1.61
I. Projected Population for year 2001	-	-	-	-	-	760	773	777	786

A SIMULATION ANALYSIS OF LAGGED FERTILITY ADJUSTMENTS IN DEVELOPING COUNTRIES TO EXOGENOUS MORTALITY DISTURBANCES

G. S. Tolley, University of Chicago
G. K. Kripalani, Western Michigan University

Several low income countries experienced substantial mortality declines in the fifties. Thereafter, mortality gains have slowed down significantly. On the fertility side, there is increasing evidence that lagged downward fertility adjustments have probably begun.

Interacting mortality and fertility disturbances have profound implications for the growth and structure of future populations in the low-income world. As mortality gains decline and fertility declines gather momentum, population growth rates in many parts of the less-developed world may tend to moderate levels, disproving the alarmist views of the prophets of gloom.

To gain approximate quantitative insights into the dynamics of these mortality fertility interactions, an initially stable population model based on suitable assumptions relevant for low income world is developed. This initially stable population is assumed to be subject to exogenous mortality disturbance at time t=o. The case discussed is that of a once-for-all mortality decline. Other assumptions are that birth rates respond in downward fashion to declines in death rates. The main elements of the hypothesis pertain to household family formation behaviour and are: (a) the concept of Desired Family Size; (b) household response to past mortality changes via lagged adjustment in planned fertility; (c) 'myopic' expectations about future mortality improvements. The expectations relation used is:

$$E\,y\,(t + c/t) = M\,(t + c/t).\ E\,y\,(t/t) \quad \ldots.(A)$$

where

$$E\,y\,(t/t) = L\,E\,y\,(t - 1/t - 1) + (1 - L)\,y\,(t - 1) \quad \ldots.(B)$$

where L = lag parameter lying between 0 and 1;
$E\,y\,(t + c/t)$ = expected change in the force of mortality in the period $(t+c)$, expectations formed at time t; and
$y\,(t - 1)$ = actual change in the force of mortality observed in the time period $(t - 1)$.
$M\,(t + c/t)$ = Myopia factor at time t for time period $(t + c)$ in the future.

The expectations hypothesis postulated is that currently held expectations (at time t) about mortality improvements per period in the future /myopia ignored/ are a weighted average of the lagged expectations held for the last period and the lagged observed value.

A female population whose family formation behaviour has a goal of achieving a fixed Desired Completed Family Size (DCFS) will respond to mortality changes by appropriate adjustment in their planned fertility. Mortality improvements unaccompanied by any downward adjustments in actual fertility will result in an accelerated population growth of existing numbers and further the households will discover that the number of children surviving to adulthood exceeds the quantity aimed at. Even if instantaneous and 'full' adjustments in planned fertility are made immediately following mortality disturbance and are realized, in the early stages for a time, however, the population will grow at a rate faster than previously on account of the fact that more females would survive to adulthood and higher ages than would have been the case in the absence of any downward disturbances in mortality. Thus mortality improvements will lead in the immediate future to an accelerated rate of population growth even if instantaneous and 'full' fertility adjustments accompany mortality changes. In cases in which fertility responses to mortality declines are neither instantaneous nor 'full' additional sources contributing to accelerated rate of population growth will operate. Both the magnitude and the duration of this process will depend principally upon the lag parameter. In elaborate models in which family formation takes place over a life and time span, the myopia parameter representing expected mortality improvements in the future will also be relevant in determining the sequence of the rates of population growth.

Formulae Derivation

A stable population subject to once-for-all mortality disturbance at time t.

Let a $(x, t+c)$ denote change in the force of mortality at age x during time period $(t+c)$. In this case, the mortality disturbance of magnitude a per period starts at time t and continues uniformly over the first time period t to t+1 and ceases at point of time t+1. Thereafter the age-specific mortality schedule at time t+1 continues

unchanged. Let u (x, t+c) refer to force of mortality at age x at time t+c. When the discussion is general and applies to all age groups, we will, for the sake of brevity use the notation u (t+c) to refer to the force of mortality at any age at time t+c. We have:

$$u(t+g) = u(t) - ga \qquad \ldots.(1)$$

where g is a fraction lying between o and 1; and

$$u(t+c+g) = u(t) - a \qquad \ldots.(2)$$

where c is a positive integer and $o<g<1$.

Let S(x, t-1) refer to before-disturbance one period survival rate schedule. When the discussion is general and applies to all age groups, we will use the notation S to refer to the pre-disturbance survival rate schedule. Let SR(t+c) refer to actual one-period survival rate for any age during time period t+c, that is from time t+c to time t+c+1. Now we have:

$$SR(t) = \exp\left[\int_0^1 u(t+g)\,dg\right]$$

$$= \exp\left[\int_0^1 [u(t) - ag]dg\right]$$

$$= S.\ \exp\lfloor a/2\rfloor \qquad \ldots.(3)$$

$$SR(t+c) = \exp\left[\int_0^1 u(t+c+g)dg\right]$$

$$= \exp\left[\int_0^1 [u(t) - a]\,dg\right]$$

$$= S.\ \exp[a] \qquad \ldots.(4)$$

for all positive integral values of c.

Let y (t) denote the change in the force of mortality at time t. We have a once-for-all decline in the force of mortality of magnitude a during period t. Hence:

$$y(t) = a$$

$$y(t+c) = 0 \quad \text{for} \quad c = 1, 2, 3, \ldots.$$

Let Ey(t+c) denote the expected periodic change in the force of mortality at time t+c. Using the distributed lag relationship, we have:

$$Ey(t) = a(1-L) \qquad \ldots.(5)$$

$$Ey(t+c) = aL^c(1-L) \qquad \ldots.(6)$$

for all positive integral values of c=1, 2, ...

Expected force of mortality at various points of time (t+c) may be written as follow:

$$E\,u(t+h) = u - a - af\,L^d\,(1-L) \qquad \ldots.(7)$$
$$d\ \ h\ \ d+1$$

where d is the integer and f the fraction part of h, for d 1 and u is force of mortality before disturbance.

The expected survival rates ESR (t+c) are given by:

$$ESR(t) = \exp\left[-\int_0^1 E\,u(t+f)\,df\right]$$

$$= S.\ \exp[a\,(1-L)/2] \qquad \ldots.(8)$$

$$ESR(t+c) = \exp\left[-\int_0^1 E\,u(t+c+f)\,df\right]$$

$$= S.\ \exp[a + a\,L^c\,(1-L)/2] \qquad \ldots.(9)$$

If D(t+c) is the planned fertility for the time period (t+c) equal to desired average number of female children born per potential mother, we have in general:

$$D(t+c).\ ESR(t+c) = DS = G \qquad \ldots.(10)$$

where S = S (t-1) before disturbance survival rate. Hence using expression for ESR (t+c), we have:

$$D(t+c) = D\exp[-a\,(1+L^c\,(1-L)/2)] \qquad \ldots.(11)$$

Let G (t+c) refer to value of long-term stable population growth factor based on period (t+c) mortality and fertility schedules. We have:

$$G(t+c) = D(t+c).\ SR(t+c) \qquad \ldots.(12)$$

Substituting the values of D (t+c) and SR we have:

$$G(t) = G\exp(aL/2) \qquad \ldots.(13)$$

since pre-disturbance long-term stable population growth rate G is equal to the product of pre-disturbance values of D and SR. In general:

$$G(t+c) = G\exp[aL^c\,(1-L)/2] \qquad \ldots.(14)$$

Thus two important propositions emerge from the two sets of values of D (t+c) and G (t+c). (1) In the periods t, t+1, ... following mortality disturbance, the survival rates are greater than the predisturbance survival rates by factor exp (a/2) for t=t and by factor exp (a) thereafter. Full fertility response would necessitate a decline in fertility by the factor exp (-a/2) during period t and exp (-a) during subsequent periods. But on account of lagged response, fertility decline factors have a sequence $\exp[-a(1-L)/2]$, $\exp[-a(1+L(1-L)/2)]$, $\exp[-a(1+L^2\,(1-L)/2)]$ ....$[\exp\ -a(1+L^c\,(1-L)/2)]$. Thus D, the average number of children born per potential mother falls gradually to the full adjustment level of D exp (-a), as c increases since L is less than unity. (2) The long-run stable population growth factor rises sharply at first from G in period (t-1) before disturbance to G exp (aL/2) in period

t immediately following once-for-all mortality gains, and then progressively declines to G exp [a L^c (1-L)/2] during period (t+c), thus asymptotically approaching the initial level of G.

SIMULATION RESULTS

An important focus of this study is to show that notions of population 'explosion' in the low-income world are unduly alarmist. Neither the very significant declines in mortality of the 1950's can continue indefinitely, much less at those high levels; nor will the fertility response to increasing actual realized family size will be too-long delayed. A detailed analyses of census data of India (not reported here) shows that there has not so far been any significant fertility response to improving mortality experience, but there is significant evidence to show that this response is in the making and underway. One evidence of this is the rate of increase in the decennial population growth rate; this rate of increase in the growth rate was substantially lower during 1961-70 decade than it was in the preceding decade 1951-60. Simulation runs based on lagged fertility response hypothesis show that will inevitably occur as fertility declines tend to shift the family size to its desired level.

One set of illustrative simulation runs has been worked out for once-for-all mortality decline case. The initial population in both cases is assumed to be a stationary population (growth rate zero or growth factor of 1.0). The once-for-all mortality decline model has a uniform mortality fall in the first period t = 0 of magnitude a = .2202 i.e. .01 per year. Abridged Life Table 1(x) column values corresponding to these assumptions are given in Annexe A Table 1. Life expectancy at birth estimates are also given in footnote (3) of the same Table. The initial t = 0 population distribution in both cases is the same and relates closely to Regional Model Life Tables population - West Females Mortality Level 7, G = 1.0 of Coale and Demeny [1] page 38.

Population projections for different values of the lag parameter L for 4 points for time t = 1, 2, 3, 4 are given in Annexe A Table 2. Important selected Characteristics of the projected populations are given in Annexe B Tables 1 through 3. Some important simulation results are discussed below.

(i) Projected populations increase over time but the rate of increase considerably slows down as time increases. This holds true for all values of L. The annual growth rate over the first period was 3.2 (per 1,000 population), increased to 5.2 in the second period, declined to 2.5 in the third period and 0.7 in the fourth period.

(ii) Projected population values were greater, greater was the value of the lag parameter.

(iii) Proportion of children in the total population which at t = 0 stood at 417 per thousand, tended to decline continuously. This was true for all L. In the FFR (Full Fertility Response Case), this proportion fell to 390 per thousand at t = 1, 352 at t = 2, 335 at t = 3 and 330 at t = 4.

(iv) The dependency ratio at first falls but then tends to rise mainly because of increasing proportions of the aged people (3+). This implies that proportion of population in potential labor force age groups rises first and then falls.

(v) Women in the child-bearing age group 1-2 increase at first but later decline; generally they stood around 30-32 percent.

(vi) Total fertility declines and asymptotically approaches the FFR value. Note over-reaction in LFR (L = .5) and NFR cases.

(viii) The long-run stable population growth factor remains at unity throughout.

ACKNOWLEDGEMENTS

Computer simulation work on this research was done by Alec E. Gores and Rodney Smith, both students at Western Michigan University. Both Mr. Gores and Mr. Smith received financial support for participating in this research from the Russell H. Seibert Fund of the Honors College of Western Michigan University.

REFERENCES & BIBLIOGRAPHY

Coale, A. J.: The Growth and Structure of Human Populations, Princeton University Press, Princeton, N. J., 1972.

Coale, A. J. and Demeny, P.: Regional Model Life Tables and Stable Populations, Princeton University Press, Princeton, 1966.

Hoover, P. F. and Longley-Cook, L. H.: Life and Other Contingencies, University Press, Cambridge, England, 1953. Published for the Institute of Actuaries and the Faculty of Actuaries.

Milbank Memorial Fund: "The Interrelations of Demographic, Economic and Social Problems in Selected Underdeveloped

Areas." New York, 1954.

Registrar General of India: Official Population Census Reports, 1951, 1961 and 1971.

Registrar General of India: Official Life Tables for India, 1941-50, 1951-60 and 1961-40.

World Bank: Population Planning-Sector Working Paper, World Bank, March 1972.

ANNEXE A TABLE 1

Abridged Life Table L(x) column Values based on Projected Mortality Disturbances

Age X	t = 0	Once-for-all Distur. t = 1, 2, 3, 4
	l(x)	l(x)
(1)	(2)	(3)
0	1000	1000
1	583	727
2	281	437
3	84	163

Note: (1) Once-for-all Mortality Disturbance Case Based on a(0) = .2202; a(i) = 0 for i = 1, 2, 3.

(2) Life expectancy at birth underlying the above data are as follows: t = 0 (14.5 Periods); once-for-all disburbance - 18.3 periods;

ANNEXE A TABLE 2

Projected Population For Different Values of Lag Parameter L. Once-For-All Mortality Improvement Case. G=1; a=.2202 (.01 per year)

Age	t=1	t=2	t=3	t=4	Age	t=1	t=2	t=3	t=4
L=0					L=.25				
0-1	417	417	417=	417	0-1	428	420	417	417
1-2	338	377	377	377	1-2	338	387	380	378
2-3	220	274	306	306	2-3	220	274	315	308
3+	94	117	146	163	3+	94	117	146	167
Total	1068	1184	1245	1262	Total	1079	1198	1237	1270
L=.4					L=.5				
0-1	435	424	420	418	0-1	440	428	422	417
1-2	338	394	384	380	1-2	338	398	387	382
2-3	220	274	320	311	2-3	220	274	323	315
3+	94	117	146	170	3+	94	117	146	172
Total	1087	1209	1270	1279	Total	1091	1217	1279	1285
L=.75					L=1.0				
0-1	452	443	436	431	0-1	466	466	466	466
1-2	338	409	401	395	1-2	337	420	420	420
2-3	220	274	332	326	2-3	220	274	342	342
3+	94	117	146	177	3+	94	117	146	182
Total	1104	1243	1315	1328	Total	1116	1277	1374	1410

Note: (1) Initial t = 0 population distribution assumed was: Age 0-1(417), Age 1-2(302), Age 2-3(197), Age 3+(84), Total (1000)

(2) G = 1 stands for growth factor of 1 per period that is, a growth rate of zero representing a stationary population at t = 0.

ANNEXE B

Projected Population and Its Selected Characteristics for Different Values of Lag Parameter L. Once-For-All Mortality Decline Case. (G=1.0, a=.2202 (.01 per year)

Characteristics	t=0	t=1	t=2	t=3	t=4
(1)	(2)	(3)	(4)	(5)	(6)
L = 0(FFR)					
1. Population Growth Rate Per Period, Per 1000	0.00	67.7	109.2	51.3	13.7
2. Annual Average Growth Rate, Per 1000	0.00	3.2	5.2	2.5	0.7
3. Proportion Children (Age Group 0-1)	.417	.390	.352	.335	.330
4. Dependency Ratio	1.002	.916	.819	.823	.848
5. Proportion Women in Child-Bearing Age Group 1-2	.302	.316	.318	.303	.299
6. Proportion Labor Force Age Groups 1-2 & 2-3	.499	.522	.550	.548	.541
7. Total Fertility [= children per potential mother]	1.38	1.23	1.11	1.11	1.11
8. Mean Age - Periods	1.45	1.49	1.58	1.65	1.67
9. Mean Age - Years	29.0	29.8	31.6	33.0	33.4
10. Life Expectancy at Birth@ - Years	29.0	36.5	36.5	36.5	36.5
11. Long Term Stable Population Growth Factor	1.00	1.00	1.00	1.00	1.00
L = 0.5					
1. Population Growth Rate Per Period, Per 1000	0.00	91.2	115.7	50.3	5.00
2. Annual Average Growth Rate, Per 1000	0.00	4.6	5.7	2.5	.25
3. Proportion Children (Age Group 0-1)	.417	.403	.352	.330	.324
4. Dependency Ratio	1.002	.958	.811	.799	.845
5. Proportion Women in Child-Bearing Age Group 1-2	.302	.309	.327	.303	.297
6. Proportion Labor Force Age Groups 1-2 & 2-3	.499	.511	.552	.556	.542
7. Total Fertility [= children per potential mother]	1.38	1.30	1.08	1.09	1.09
8. Mean Age - Periods	1.45	1.47	1.57	1.65	1.69
9. Mean Age - Years	29.0	29.4	31.4	33.0	33.8
10. Life Expectancy at Birth@ - Years	29.0	36.5	36.5	36.5	36.5
11. Long Term Stable Population Growth Factor	1.00	1.06	.97	1.00	1.00
L = 1.0 (NFR)					
1. Population Growth Rate Per Period, Per 1000	0.00	116.0	114.4	107.6	102.6
2. Annual Average Growth Rate, Per 1000	0.00	5.5	5.4	5.1	4.9
3. Proportion Children (Age Group 0-1)	.417	.417	.365	.339	.331
4. Dependency Ratio	1.002	1.002	.840	.803	.850
5. Proportion Women in Child-Bearing Age Group 1-2	.302	.302	.329	.306	.298
6. Proportion Labor Force Age Groups 1-2 & 2-3	.499	.499	.554	.555	.540
7. Total Fertility [= children per potential mother]	1.38	1.38	1.11	1.11	1.11
8. Mean Age - Periods	1.45	1.45	1.53	1.62	1.67
9. Mean Age - Years	29.0	29.0	30.6	32.4	33.4
10. Life Expectancy at Birth@ - Years	29.0	36.5	36.5	36.5	36.5
11. Long Term Stable Population Growth Factor	1.00	1.25	1.00	1.00	1.00

DISCUSSION

Che-Fu Lee, The Catholic University of America

All of the five contributed papers just presented deal more directly with issues of assessing differences and changes in "vital rates" of some kind and less directly with estimates of such demographic rates. No one addresses problems of paucity of basic demographic data, which have in the past been the preoccupation of demographer-statisticians studying the demography of developing countries. If the subjects chosen by these contributors serve as an index, it perhaps signifies the recent improvement of the data situation in most of today's developing countries. Moreover, the subjects presented in these papers are of general methodological interest in demographic analysis aside from their ramification for studying the "vital rates in developing countries", as the title of this session seems to delimit. I shall now comment on each of these fine studies in their turn.

1. The title of Udry et.al.'s paper, "Random Variation in Rates Based on Total Enumeration of Events", suggests in effect three aspects of the problem in constructing a rate: a) ascertaining the random process underlying the occurrence of an event, b) identifying the appropriate population "at risk" of the event occurrence for the denominator, and c) total enumeration of the numerator, observed occurrences. Ideally, the latter two problems should have been resolved when a model random (or stochastic) process is conceptualized, and data are collected accordingly. To estimate a rate in practice, however, one often has to rely on the available records. Socially significant events like births, marriages, and deaths or accidents tend to be registered as they occur and thus are the result of a complete count. The denominator population "at risk" on the other hand may be obtained from a different source, which is not infrequently based on a sample estimate. An even-handed treatment of sampling fluctuation in the denominator and errors involved in enumeration of the numerator, which is supposedly subject to no sampling error, is itself a difficult task. Udry et.al. did not focus their study on this issue and tacitly assumed no sampling error for a computed rate when the numerator represents a total count.

The paper moved directly into a demonstration that observed variance of crude birth rates exceeds the variance that can be expected from a binomial model, even after the variations across comparative units and over time periods have been taken into account. The authors went on to suggest other sources of non-sampling variations: correlations between random error of birth rates and time; and unequal risks of birth among individuals, i.e., heterogeneity in the probability process or compound probability distributions. These, in other words, are equivalent to saying that a simple binomial model is inadequate for depicting the "true" process of birth. This is hardly a surprise to those researchers who are inclined to model building. However, the well organized exposition in this article serves its purpose in calling the practitioners' attention to an unconservative inference on difference or change in birth rates at their face values.

One may well ask a logical question: so what should be done then? The authors seemed to suggest two ways. One is to set a minimally required sample size ensuring stability of rate estimates; the other, as the authors put it, "predicted variances made from detailed data on the actual population being studied". More concrete suggestions than these open-ended ones may require more work than the scope of the paper intended by the authors. It will suffice to point out that these suggested directions for tackling the remaining problem may be more complex than they appear at first blush.

One of such difficulties was touched by the authors in their statement, "assuming that we are usually dealing with populations in which p (the rate) has some unknown distribution, our predicted variances based on the simple binomial model seem doomed to be overestimates" (pp. 15-16). The implications of this were not pursued. Let me extend it as a query. As we conceive of the random process in terms of a more realistic and usually more complex model, the larger will become the predicted variance due to random variation. While comparing to a simple binomial assumption we tend to draw non-conservative assertions of true differences in rate. Wouldn't any refined conceptual model quickly "step up" the predicted variance and render us a "too conservative" inference, as observed variations in rates hardly ever exceed the predicted variations based on a complex model? This strikes me as a major caveat in most of the model-building exercises, e.g., birth-interval models considering the elements of fecundability, postpartum lapsed period and various outcomes of a pregnancy, etc. The most common fate of an elegantly constructed model is being shelved and never becoming useful in data confrontation, especially for detecting differences or changes in demographic rates.

The alternative to a preconceived model depicting the random process lies

in data exploratory-confirmatory approach (Tukey 1970). The authors' call for studying the detailed data on the actual population may be interpreted as suggesting this line of approach, but I am not sure from reading their paper. Exploration of data distributions and boundaries of homogeneity and heterogeneity requires detailed information on differentials in rates. A pragmatic approach without having to specify the underlying random (probability) process in the first place is suggested by Allen and Avery; this leads our discussion to the next paper.

2. If a sample is drawn from the population at risk of birth, the binomial distribution of births and no births or a multinomial distribution by number of births during a period of observation can be handled as a discretely measured dependent variable, without being converted into rate measurement; and such a frequency distribution can be cross-tabulated with other categorical factors in a multiple-way contingency table and analyzed by a log-linear analysis for significant factors or interactions among them in differentiating the distributions of the observed frequency of event occurrence. This is exactly what Allen and Avery proposed in their paper.

It seems to me a promising new way of analyzing differentials in demographic measures by discretizing the occurrence of events. Allen and Avery treated the period fertility in terms of frequencies of mothers falling into a dichotomy of having had no birth and one or more births during the past five years before the Costa Rican census. This dependent variable was alternatively measured in terms of a polytomy of 0 to 5+ births. The odds of falling into one category vs. the other(s) is actually the criterion quantity to be analyzed. Since all the predicting factors selected (rural-urban residence, marital status, labor force status, and education), and the control variables (previous parity and age) are all represented as categorical measures, the log-linear model for discrete multivariate analysis seems to be suitable. The analytical results were then presented in terms of variations in odds of having any births vs. no births (or having one particular number of births vs. all others in a polytomous measure of the dependent variable), which are attributable to differences in the selected predicting factors and their interaction effects, which have been identified as statistically significant.

Judicious presentation of the statistics resulting from the log-linear analysis is essential in making important findings recognizable, as Davis (1975) once complained that such an analysis generated "too many results". Allen and Avery presented at length the variations in odds by various-ordered effects. Such odds figures filled almost six full pages of table presentation, and their graphs attempted for facilitating a visual summary of the results did not seem to alleviate much of the reader's burden in putting these tremendous numbers of odds figures into perspective. Showing the possibility of constructing the probability of birth (convertible into the familiar birth rate) from the odds figures in the appendix, the authors unfortunately failed to see the importance of presenting the "smoothed" fertility rates as differentiable by the tested factors. I am inclined to think of discrete multivariate statistics as the means and the vital rates arrivable through statistical testing and smoothing being the end. I am sure that Allen and Avery can easily produce the familiar differential birth rates following their log-linear analysis if they elected to do so. It would involve use of the model predicted frequencies, rather than the observed frequencies, and computing the rate thereof.

Just a point of information: the odds measure in the case of a dichotomous variable like mortality - death or no death, is easily interpretable. The odds measure in the case of polytomy is limited in the log-linear analysis to the odds of a chosen category to all others. This sometimes may not be easy to interpret. There is at least one alternative method for polytomous dependent variables like fertility observed over a longer period. The weighted least square approach (Grizzle et.al. 1969) to multivariate analysis of categorical data allows for the flexibility of converting a polytomous dependent variable into its expectation, i.e., the birth rate in this case. Moreover, the predicting factors are not limited to categorical measures, and the hypotheses need not be hierarchical in the weighted least square method.

3. Rashid and McElroy's comparison of the labor force separation rate for Saudi Arabia obtained through a longitudinal study and that obtained through standard working life table, seemed to have raised more questions than it answered. My first expectation from the title of this paper was in seeing a discrepancy which often results in comparisons between period and cohort rates: one being cross-sectional rates for different age groups synthesized, and the other tracing the flow through ages of an actual cohort. However, the so-called "longitudinal study" in this paper refers to a two time observation apart only by a period of 9 months. Without a detailed explanation of the computation procedures in the paper, I am at a loss in finding justification for calling such a short period data "longitudinal". The period withdrawl rates for an occupation-age category group were not clearly explained either. I was puzzled as to whether or

not the "longitudinal" meant a prospective measure comparing the job status at the end of 9 months subsequent to the beginning of the survey; and whether the rate used for working life table analysis was a retrospective job status last year compared with that at the beginning of the survey. The results of 182% difference in withdrawl rates obtained from the life table and the longitudinal data in professional, technical, and managerial category, and 60% in production workers, operatives and laborers, were indeed alarming as expressed by the authors, but no adequate explanation for these discrepancies were given. Could they be due to the current age structure of the occupational make-up: modern sector occupations are filled by younger males (e.g., the professional), so that the relatively small proportions of higher aged males overrepresent the withdrawl rate from one age level to another, in a cross-sectional comparison? Questions like these must be answered by the authors in a fuller presentation of their computational details.

4. The two papers by Kripalani and his associates on the model of population growth may be discussed together. Their simulation analysis reminded me of Frejka's (1973) work entitled "The Future of Population Growth: Alternative Paths to Equilibrium". Frejka used vital statistics available around 1965-1970 and projected the population growth to the year 2150 following alternative assumptions of reaching an equilibrium (just replacement rate, NRR = 1) immediately, in 10, 20 years or a longer period. The major innovation here is to take the initial rate of growth, instead of NRR=1, as a point of reference. It also quantifies the "lagged" response of fertility decline to the initial "disturbance" of reduced mortality by a parameter between 0 and 1; the immediate fertility reduction to offset the effect of declined mortality at one end, and no fertility response hence allowing for the full effect of the initial mortality change on the growth of population, at the other. Of course, there are finer manipulations of the input variables in the present simulation analysis than the abstract linear adjustment of fertility and mortality schedules in the population projection as conducted by Frejka. Like other well conceived projection exercises, Kripalani and his associates have added to the material that is useful for population education needed by decision makers and development planners, who are concerned with the dynamics of population growth and want to be told about differences in terms of quantitative magnitude.

It may be interesting to note that Kripalani and Smith's projections of the Indian population based on alternative assumptions of lagged fertility response ("L" ranging from 0.6 to 1.0) fall into a rather small range of variation, compared to the alternative projections carried out by Frejka, who assumed various lengths for the lapsed period before NRR= 1 is reached. In terms of the projected total population in the year 2000, the four projections of Kripalani and Smith come very close to Frejka's project no. 2, which assumes that a just replacement fertility rate is attained in the years 1980-1985. I admit that there are technical problems involved in such a comparison across projections done by different demographers who all have their respective justifications in generating the projected figures of their own. What I fear is: can we expect non-demographers to understand our projection exercises, or simply tell them to make their own choice according to their own taste.

REFERENCES

Davis, James A. 1975. "Key Concepts in 'The Goodman System' for Analysing Contingency Tables: An Outline" read at Meetings of the American Sociological Association, San Francisco, Calif. August 1975

Frejka, Tomas. 1973. The Future of Population Growth: Alternative Paths to Equilibrium. New York: John Wiley.

Grizzle, J. E., C. F. Stramer and G. C. Koch. 1969. "Analysis of Categorical Data by Linear Models" Biometrics 25(Sept.):489-504.

Tukey, J. W. 1970. Exploratory Data Analysis. Reading, Mass.: Addison-Wesley.

DESCRIPTION OF THE SURVEY OF INCOME AND EDUCATION (SIE) OPERATIONS

George H. Gray, U.S. Bureau of the Census
Marvin M. Thompson, U.S. Bureau of the Census

INTRODUCTION

The Office of Education of the Department of Health, Education, and Welfare (HEW) has for a number of years distributed funds authorized by Title I of the Elementary and Secondary Education Act of 1965, utilizing a formula that includes the estimate of the number of children 5 to 17 years of age in poverty families in each State. Since 1972, the estimate used has been the number of poor children in 1969, according to the 1970 Census of Population and Housing. As we move further in time from the census, the interstate relationships for children in poverty are likely to be changed because of changes in population growth, family formation and dissolution, and economic activity. Since 1970, national estimates of children in poverty have been available from the Current Population Survey (CPS). However, CPS estimates were not sufficiently reliable on a State basis to substitute for the census figures.

Accordingly, Congress in enacting the Educational Amendments of 1974 (Public Law 93-380) provided in section 822(a) that, "The Secretary of Commerce shall, in consultation with the Secretary of Health, Education, and Welfare, expand the current population survey (or make such other survey) in order to furnish current data for each State with respect to the total number of school-age children in each State to be counted for purposes of section 103(c)(1)(A) of title I of the Elementary and Secondary Act of 1965." Pursuant to this legislative requirement, the Bureau of the Census in cooperation with agencies of HEW, mounted the Survey of Income and Education (SIE) and carried it out between April and July 1976 at a sample of approximately 190,000 designated addresses.

The SIE was also designed to satisfy the requirements of section 731(c)(1) of the Bilingual Education Act, Title VII, ESEA as amended by Public Law 93-380, which authorizes the Commissioner of Education to estimate from a survey the number of children and other persons in the States who, because of limited English-speaking ability, are in need of bilingual education, guidance, and counseling.

Finally, at HEW's request, the opportunity presented by such a large survey was used to gather some additional income-related information such as receipt of food stamps, housing costs for homeowners and renters, and estimated cash assets. Also, information relevant to a number of HEW programs was collected, including data on education, disability, health insurance coverage, and institutionalized persons.

SURVEY DESIGN

The primary objective of the Survey of Income and Education was to determine for each State the number of children 5-17 years of age in poverty. In discussions with HEW and the Congressional staffs involved, it was agreed that the criterion to be used for providing equity among the States was an estimated coefficient of variation (C.V.) of 10 percent for the count of poverty children. A preliminary sample design was created to yield this reliability for that statistic. Since we were also interested in obtaining reliable estimates of persons with limited English-speaking ability, additional cases had to be added. While we were able to achieve an estimated coefficient of variation of 10 percent or better on persons with limited English-speaking ability for most States, the estimated C.V. for 12 States was above this level and ranged up to an estimated C.V. of 20 percent.

SAMPLE DESIGN

The sample was designed to be State representative and was to be completely independent of other Census samples, such as the CPS. The sample for SIE was a stratified multistage non-compact cluster design. For the first stage of selection, each State was divided into areas called Primary Sampling Units (PSU's). These areas were either a Standard Metropolitan Statistical Area (SMSA) or a group of geographically-neighboring counties or independent cities. The PSU's were then grouped in strata based on estimates of like characteristics derived from the 1970 Census. The primary determination of strata classification was the proportion of children 5-17 years of age living in poverty, based on 1970 Census data. PSU's with large populations in relation to the sampling rate for the State formed strata by themselves and came into sample with certainty.

In eight States (Connecticut, Delaware, Hawaii, Maryland, Massachusetts, New Hampshire, Rhode Island and Vermont) and the District of Columbia, every PSU was selected for sample with certainty. In the remaining States, two PSU's were selected from each strata that were not large enough to be in sample with certainty.

Within each PSU, the majority of the sample of housing units and group quarters were selected from the list of units in the 20-percent sample of the 1970 Census. The 20-percent sample file was used because it provides the information on income and poverty which determined the stratification of the sample.

In order to represent persons living in units completed since the 1970 Census, a sample was selected from the building permits issued since 1970 in those areas under the jurisdiction of building permit offices. This represents the majority of this type of unit. For the remaining areas (those without a building permit office), a sample of units built since 1970 was obtained by

selecting such units in the area segments from recently-retired CPS samples.

Finally, the SIE sample included units selected from (1) a list of special places, such as rooming and boarding houses, communes, flop houses, military installations (excluding military barracks), agricultural workers' dormitories, etc., and (2) a list of mobile homes in mobile home parks established since the 1970 Census.

QUESTIONNAIRE CONTENT

Public Law 93-380 amends section 103 of the Secondary and Elementary Education Act of 1976 to read, "... in determining the families which are below the poverty level, the Commissioner (of Education) shall utilize the criteria of poverty used by the Bureau of the Census in compiling the 1970 decennial census." In the years since 1970, the same definition has been used in the Current Population Survey's March Income Supplement to determine poverty status though it is updated annually to reflect changes in the Consumer Price Index. As previously noted, section 822(a) of Public Law 93-380 specifically mentions expansion of the current population survey as an acceptable method of determining the number of poor children. In addition, the existence of a processing system based on CPS made it possible to meet the stringent deadlines imposed by the Congressional mandate. Finally, very serious consideration was being given to combining the SIE and CPS to provide a larger sample for estimates of the count of poor children. For these reasons, it was decided that SIE would replicate exactly the March CPS questionnaire content though it would be expanded to cover additional subject matter. Therefore, the core questions on current labor force status, last year's work experience and money income, together with such demographic variables as age, sex, marital status, family membership, household membership, veteran status, educational attainment and ethnic origin, are asked and recorded in the same manner on both questionnaires.

The items on foreign birth, language or languages spoken in the household and language spoken in the home when the sample person was a child are screening questions to determine if the questions on English Language Proficiency should be asked. These last questions (what language the sample person speaks, how well the person speaks and understands English, what language does he usually speak to friends, and what language does he usually speak to his children, or in the case of children speaks to his brothers or sisters) are used as a Measure of English Language Proficiency (MELP). This series was developed by the Center for Applied Linguistics under a contract with the National Center for Education Statistics.

For the foreign born, there are questions to determine when they came to the United States to stay and where they were born. All sample persons are asked how long they have lived in the State and, for movers, what State they lived in before moving to the State of residence at the time of interview. These questions will be used to develop measures of immigration to the United States and measures of internal migration.

Additional items asked for all persons, though screened on appropriate age groups, include questions on school enrollment, disabilities that limit the person's ability to attend school, limit or keep the person from working at a job, or limit the amount or kind of housework they can do. For those with a limitation, it is determined how severe that limitation is, the cause of the limitation and who diagnosed it. Finally, for each person, questions are asked concerning their coverage by health insurance plans or other programs that provide health benefits or services and whether they received any of these benefits or services in the past year.

For the household as a whole, information was collected on food stamp recipiency in 1975 and 1976, cash assets, mortgage or rent payments and if a rental unit, whether or not it was subsidized. While the data from these questions and those on education, disability and health insurance will be used to meet the needs of various programs sponsored by HEW, they are more specifically to be used in the estimation of costs and caseloads under various alternative assumptions about eligibility for programs such as food stamps and AFDC. In addition, they will be used to analyze the impact of the inclusion of such in-kind costs and assets on alternative definitions of poverty.

Finally, there are a set of questions designed to determine household membership during the reference year (1975). These questions will be used in research concerning the effect of changing household membership on the income and size of the family and hence on their poverty status.

DATA COLLECTION

Interviewing was begun in late April 1976 and extended through July of that year. Approximately 95 percent of the workload was completed during the months of May and June.

The 191,459 assigned households were located in approximately 1,800 counties and independent cities. To complete this task required 2,500 interviewers, of whom 1,600 or 63 percent were new to Census operations. About one-fifth of the interviewers had worked on CPS (including March 1976) and the remainder were working on other Census surveys at the time they were assigned to SIE. In addition, about 200 persons were hired as crew leaders whose primary function was as reinterview specialists though they performed other tasks such as aiding new interviewers begin their work and observing and helping those who needed additional training. The crew leaders also assisted in reducing the number of refusals and other noninterviews, especially in areas with high noninterview rates. The data collection effort was coordinated through the Bureau's 12 regional offices, where the regular staff was supplemented by supervisory and clerical help to perform the extensive reviewing of the questionnaires required.

Interviewers and office clerks completed a 4-hour home-study which introduced them to the survey and

the forms to be used. They were then given 3 days of classroom training on the concepts to be applied and procedures to be used in interviewing. During their training they were led through several practice interviews to familiarize them with the content and skip patterns on the questionnaire. Following this, they completed a 6-hour post-classroom home-study which gave additional training and tested them on the training already received. All newly-hired interviewers were given 2 days of on-the-job training during which they were accompanied by more-experienced personnel who demonstrated interviewing techniques and observed them perform several interviews before leaving them on their own.

Interviewing for SIE was conducted by personal visit to the assigned address. Any responsible adult, that is someone who was knowledgeable about the work patterns and income of the family, could act as the respondent for the entire household. While technically, anyone over 14 years of age could be a household respondent, in practice, teenagers were accepted only as a last resort. In most cases, the respondents were the head of the household or the head's spouse. In any case, the interviewers were encouraged to make extensive callbacks either by phone or personal visit to obtain more precise information when not available from the household respondent. While the average time required to complete an interview was about 45 minutes, some households took much longer, especially when callbacks were required.

QUALITY CONTROL MEASURES

Throughout the period of interviewing, the questionnaires received from the interviewers were closely monitored to determine the number and type of noninterviews each was reporting. If the number seemed excessive, a crew leader or supervisor contacted the interviewer to explore the problem and help reduce the noninterviews. During the latter part of June and July, weekly reports were made by the Regional Offices, setting forth the noninterviews by type for each State in their regional area. A target of 5 percent for noninterviews at occupied households and 20 percent for all types was set for each State. During July, crews of experienced interviewers and supervisors visited those States with rates above the targets and attempted to reduce them. While their efforts met with considerable success in most places, a few States remained above the target noninterview rates when field work was closed out.

Nationwide, the noninterview rate for occupied households was 4.6 percent, which is identical to the like rate for CPS in April, May and June of 1976. The noninterview rate for all types of assigned addresses (occupied, vacant, demolished, condemned, etc.) was 21.0 percent, which compares to 20.3 percent for CPS.

While the noninterview rate for occupied households exceeded 5 percent in 15 States, this rate exceeded 6 percent in only 5. The highest rates were posted in the District of Columbia--13.5 percent, Alaska--8.1 percent, and Nevada--7.5 percent. On the other hand, 11 States recorded noninterview rates for occupied households below 3 percent. The lowest was 1.7 percent for Arkansas.

The other major control on the quality of interviewing was an extensive and detailed review of the questionnaires as they were turned in by the interviewers. The first of these was at the Regional Office level at which time the first 25 of the questionnaires returned by the interviewer were reviewed in their entirety. If certain critical items were mishandled or left blank, the Regional Office contacted the interviewer to correct the error or directly called the respondent for missing information. As the family income is an important determinant of poverty status, the Regional Office continued to review the income items on all remaining questionnaires beyond the first 25 from each interviewer and continued to try to obtain any information missing in that area.

When the questionnaires arrived in the Census Bureau's Processing Center in Jeffersonville, Indiana, the review of the questionnaires was repeated; this time on every item on every questionnaire. The Regional Offices were notified of any systematic errors. While it was now too late to call upon the respondent for missing information, the Regional Office could contact the interviewer and correct the problem for the remaining interviews. The single greatest gain from the Processing Center review was the correct marking of the machine-readable data and numbers on the questionnaire. For example, in the income area both a write-in space and machine-readable numbers are used to record each response. The most common interviewer error was failure to mark the machine-readable numbers. On items for which records were kept, this type of error was reduced from 1 percent of the entries to .3 percent for any one item. In light of the large number of newly-hired interviewers, this compares favorably with the .2 percent blank rate per item for the Current Population Survey.

In addition to the close supervision of interviewers and extensive review of the questionnaires, two other procedures were used to control the quality of the interviewing. The first was a telephone recheck of the interviewer's work. The rechecker verified with the interviewed household the list of household members and then re-asked five items that pertained to the household as a whole and five items that had been asked for each household member. The recheck responses were compared to the original and differences were reconciled. Any differences attributable to the original interviewer were discussed with him and remedial training provided where necessary. The first three interviewed households returned by the interviewer were rechecked and thereafter, one interview was rechecked every 2 weeks the interviewer continued working. On the average, seven interviews were checked out of the total workload of approximately 80 assigned addresses per interviewer.

The second procedure used as a quality control was the reinterview of a 5-percent systematic sample of the households assigned. The reinterview was conducted by a staff of interviewers who were more

thoroughly trained than the average SIE interviewer and which had a higher proportion of interviewers from CPS and other Bureau programs. While the questionnaire used by the reinterviewer differed markedly from that used in the original interview and there was a time lag between the two visits to the address necessitated by the sampling procedure, nevertheless, the reinterview did uncover some gross errors on the part of the interviewers and these were fed back through the Regional Office staff. The reinterview, together with the check for units missed in the Census is, of course, far more important as part of the overall evaluation of the quality of the data than as an interviewer control.

PROCESSING THE DATA

After reviewing the questionnaires for errors and correcting those for which the information was available, the Processing Center personnel entered codes for all industry and occupation responses, grouped all members of the household into families according to their relationship to the head of the household and entered codes where necessary, and where appropriate, coded State of previous residence for movers.

All clerical reviewing and coding was rechecked on a 100-percent basis to assure an acceptable level of quality. The clerical review, coding and verification operations took place during June, July and August of 1976 and required the services of approximately 40 persons working full time. Approximately 160 mandays of overtime were also required to meet the deadline.

The SIE questionnaires were then microfilmed and the data transferred to computer tape by means of the FOSDIC process. FOSDIC (Film Optical Sensing Device for Input to Computers) is a programmable machine that scans the developed film to ascertain the presence or absence of a mark in the coded dots or numeric figures on the questionnaire and transfers this information to a magnetic computer tape.

This computer tape is then run through the computer and processed by a Data Acceptance Program. The Data Acceptance Program checks the filming operation to assure that all required pages have been filmed and that index marks used in the FOSDIC program have been properly recorded. It also verifies that certain critical data have been correctly entered, such as the Household Identification Number, the Interview/Noninterview status of the household and if noninterview, the type of noninterview. If a questionnaire fails one of these or any of the other checks in the Data Acceptance Program, it is rejected. The error then has to be corrected, the questionnaire refilmed and recycled through the Data Acceptance Program. Most questionnaires are accepted on the first pass. However, approximately 11,700 or 5.5 percent of the SIE questionnaires were recycled, a few for more than one time. After all questionnaires had been through the Data Acceptance Program at least once, some 75 were dropped from the file as they were rejected again and time had run out in late October.

The accepted records were then passed through a series of programs to edit the labor force, work experience in 1975 and income questions. Those programs were the same as used in producing the March 1976 CPS file. They not only edit the data but create a number of recodes used in tabulations, and impute missing data, including income. These programs were used to produce, as closely as possible, data that would fulfill the Congressional requirement to use the same poverty definition as was used for the 1970 Census.

The remaining data on the SIE questionnaire were subjected to a consistency edit that made only those changes that could be inferred from the data themselves and did not impute for any missing information.

Each stage of the editing and imputation programs provided for printouts of actual data or counts of changes so that the operations could be reviewed.

After the editing and imputation had been reviewed and the file accepted, it was then weighted to represent the population as a whole. Initially, each record was assigned a base weight that was the reciprocal of the probability of sample selection. Next, factors were applied to adjust for occupied households that were not interviewed. Adjustments were then made to account for differences between the sample areas chosen and the strata from which they came. The resultant weights were summed and compared to independent estimates of the national population in 116 age-sex-color categories. Factors derived from this comparison were then applied to the individual weights. Finally, the weights were again summed and factors applied for three age groups (5-17 years old, 65 years old or older, and all other ages) for each State and the District of Columbia. To bring these last two groups of estimates into closer agreement, the adjustments were iterated a total of three times.

During the weighting process, factors for the national age-sex-color controls are calculated and at that time ratios of the coverage of the population in those various groups are produced. These revealed that SIE had a coverage ratio of 93 percent as opposed to 96 percent for CPS. This is in addition to the undercoverage experienced by the 1970 Census, as the independent estimates of population used as controls for both surveys are derived by updating the census counts by taking into account births, deaths and migration since that time. SIE had coverage of the population equal to or better than CPS for Blacks and other races. It was appreciably lower for Whites, both males and females in almost every age category.

TABULATIONS AND TAPE FILES

Counts of the children 5-17 years of age in poverty for each State were produced in December 1976 and after a review and analysis of these data, a preliminary report was forwarded to Congress on February 18, 1977. A final report incorporating the results of the evaluation is expected to be sent forward in October 1977.

Tabulations have been produced and forwarded to various groups at HEW, Department of Labor, and Census, covering food stamp recipiency, public assistance, child care and labor force status of mothers, characteristics of families and unrelated individuals, income, characteristics of persons with language difficulties, school enrollment, the educationally handicapped, health insurance, work experience in 1975 and labor force status for a number of geographical areas.

Computer tapes have been provided to HEW, the Congressional Budget Office, and the Civil Rights Commission Age Discrimination Study to aid in analyzing the impact of various alternative changes to the welfare system.

A tape is being prepared for general public use that will carry all the information collected by the SIE. All 50 States and the District of Columbia will be identified on the tape. In addition, 122 SMSA's will be identified and within the limits of the Bureau's confidentiality restrictions, the central city of the SMSA, the remainder of the SMSA and the nonmetropolitan areas of the State. The tape will contain individual records for 336,405 persons 14 years old or older, including 2,769 members of the Armed Forces and records for 104,410 children 0-13 years of age. There are summary records for 160,973 families or unrelated individuals as well as 151,170 records for the interviewed households. The tapes and information concerning them can be obtained from:

Customer Services Branch
Data User Services Division
Bureau of the Census
Washington, D.C. 20233

SIE NONINTERVIEW RATES BY STATE

	1. Total Hhlds	2. Interviewed Hhlds	3. Occupied Hhlds (2+4)	Type A Nonint. 4. Number	Type A Nonint. 5. Rate (4÷3)	Type B Nonint. 6. Number	Type B Nonint. 7. Rate (6÷1)	Type C Nonint. 8. Number	Type C Nonint. 9. Rate (8÷1)	Type A+B+C NI's 10. Number (4+6+8)	Type A+B+C NI's 11. Rate (10÷1)
UNITED STATES	191,459	151,170	158,475	7,305	4.6	24,600	12.8	8,384	4.4	40,289	21.0
NEW ENGLAND:	26,970	20,754	21,604	850	3.9	4,501	16.7	865	3.2	6,216	23.0
Maine	3,123	2,189	2,240	51	2.3	734	23.5	149	4.8	934	29.9
New Hampshire	5,884	4,261	4,434	173	3.9	1,265	21.5	185	3.1	1,623	27.6
Vermont	3,752	2,723	2,796	73	2.6	822	21.9	134	3.6	1,029	27.4
Massachusetts	4,614	3,664	3,879	215	5.5	616	13.4	119	2.6	950	20.6
Rhode Island	4,193	3,386	3,509	123	3.5	546	13.0	138	3.3	807	19.2
Connecticut	5,404	4,531	4,746	215	4.5	518	9.6	140	2.6	873	16.2
MIDDLE ATLANTIC:	16,506	13,459	14,323	864	6.0	1,662	10.1	521	3.2	3,047	18.5
New York	5,276	4,211	4,521	310	6.9	585	11.1	170	3.2	1,065	20.2
New Jersey	5,684	4,694	5,007	313	6.3	518	9.1	159	2.8	990	17.4
Pennsylvania	5,546	4,554	4,795	241	5.0	559	10.1	192	3.5	992	17.9
EAST NORTH CENTRAL:	25,797	20,933	21,905	972	4.4	2,913	11.3	979	3.8	4,864	18.9
Ohio	5,508	4,501	4,766	265	5.6	558	10.1	184	3.3	1,007	18.3
Indiana	4,820	3,965	4,083	118	2.9	550	11.4	187	3.9	855	17.7
Illinois	5,480	4,499	4,776	277	5.8	474	8.6	230	4.2	981	17.9
Michigan	5,744	4,450	4,669	219	4.7	810	14.1	265	4.6	1,294	22.5
Wisconsin	4,245	3,518	3,611	93	2.6	521	12.3	113	2.7	727	17.1
WEST NORTH CENTRAL:	25,592	20,448	21,230	782	3.7	3,198	12.5	1,164	4.5	5,144	20.1
Minnesota	4,238	3,485	3,579	94	2.6	496	11.7	163	3.8	753	17.8
Iowa	4,694	3,879	4,000	121	3.0	479	10.2	215	4.6	815	17.4
Missouri	3,088	2,343	2,463	120	4.9	450	14.6	175	5.7	745	24.1
North Dakota	3,644	2,922	3,007	85	2.8	493	13.5	144	4.0	722	19.8
South Dakota	2,365	1,765	1,846	81	4.4	371	15.7	148	6.3	600	25.4
Nebraska	3,624	2,932	3,075	143	4.7	427	11.8	122	3.4	692	19.1
Kansas	3,939	3,122	3,260	138	4.2	482	12.2	197	5.0	817	20.7
SOUTH ATLANTIC:	22,052	17,098	18,031	933	5.2	3,042	13.8	979	4.4	4,954	22.5
Delaware	3,001	2,310	2,455	145	5.9	444	14.8	102	3.4	691	23.0
Maryland	3,262	2,714	2,869	155	5.4	326	10.0	67	2.1	548	16.8
Dist. of Columbia	2,172	1,578	1,824	246	13.5	249	11.5	99	4.6	594	27.3
Virginia	2,478	2,036	2,122	86	4.1	238	9.6	118	4.8	442	17.8
West Virginia	2,073	1,671	1,709	38	2.2	234	11.3	130	6.3	402	19.4
North Carolina	1,997	1,555	1,613	58	3.6	310	15.5	74	3.7	442	22.1
South Carolina	1,895	1,380	1,441	61	4.2	323	17.0	131	6.9	515	27.2
Georgia	1,937	1,534	1,582	48	3.0	242	12.5	113	5.8	403	20.8
Florida	3,237	2,320	2,416	96	4.0	676	20.9	145	4.5	917	28.3
EAST SOUTH CENTRAL:	8,057	6,361	6,552	191	2.9	982	12.2	523	6.5	1,696	21.1
Kentucky	1,970	1,517	1,587	70	4.4	275	14.0	108	5.5	453	23.0
Tennessee	2,185	1,736	1,791	55	3.1	253	11.6	141	6.5	449	20.5
Alabama	2,055	1,653	1,686	33	2.0	231	11.2	138	6.7	402	19.6
Mississippi	1,847	1,455	1,488	33	2.2	223	12.1	136	7.4	392	21.2
WEST SOUTH CENTRAL:	11,531	9,158	9,511	353	3.7	1,357	11.8	663	5.7	2,373	20.6
Arkansas	1,925	1,505	1,531	26	1.7	259	13.5	135	7.0	420	21.8
Louisiana	2,065	1,659	1,735	76	4.4	196	9.5	134	6.5	406	19.7
Oklahoma	2,429	1,896	1,989	93	4.7	287	11.8	153	6.3	533	21.9
Texas	5,112	4,098	4,256	158	3.7	615	12.0	241	4.7	1,014	19.8
MOUNTAIN:	33,755	26,383	27,773	1,390	5.0	4,447	13.2	1,535	4.5	7,372	21.8
Montana	3,963	3,034	3,190	156	4.9	538	13.6	235	5.9	929	23.4
Idaho	5,879	4,568	4,773	205	4.3	843	14.3	263	4.5	1,311	22.3
Wyoming	4,536	3,569	3,741	172	4.6	565	12.5	230	5.1	967	21.3
Colorado	3,782	3,014	3,174	160	5.0	478	12.6	130	3.4	768	20.3
New Mexico	2,589	2,077	2,164	87	4.0	307	11.9	118	4.6	512	19.8
Arizona	2,705	2,042	2,160	118	5.5	447	16.5	98	3.6	663	24.5
Utah	5,110	4,136	4,309	173	4.0	616	12.1	185	3.6	974	19.1
Nevada	5,191	3,943	4,262	319	7.5	653	12.6	276	5.3	1,248	24.0
PACIFIC:	21,199	16,576	17,546	970	5.5	2,498	11.8	1,155	5.4	4,623	21.8
Washington	4,406	3,567	3,743	176	4.7	487	11.1	176	4.0	839	19.0
Oregon	4,841	3,944	4,141	197	4.8	486	10.0	214	4.4	897	18.5
California	5,067	4,202	4,432	230	5.2	465	9.2	170	3.4	865	17.1
Alaska	3,677	2,360	2,568	208	8.1	668	18.2	441	12.0	1,317	35.8
Hawaii	3,208	2,503	2,662	159	6.0	392	12.2	154	4.8	705	22.0

EVALUATION OF INCOME REPORTING ON THE SURVEY OF INCOME AND EDUCATION

John Coder, Bureau of the Census

INTRODUCTION

The collection of income data in household surveys is one of the most difficult tasks for the Bureau of the Census. Nonresponse rates to questions concerning income on Census Bureau surveys have traditionally been higher than nonresponse rates for any other subject matter. Not only is nonresponse a serious problem, research has also shown that responses to the income questions have significant errors of reporting amounts and reporting of no amount when an amount was actually received. Because accurate income information is difficult to obtain and because the Survey of Income and Education (SIE) had as its major objective to measure the number of poor school age children in each State for the purposes of equitably distributing Federal educational funds, it was particularly important to evaluate the accuracy of the income data collected in the SIE.

This paper deals with several aspects of the evaluation of SIE income statistics. These aspects include: 1) a comparison of SIE data collection and processing techniques with those of the March Current Population Survey, 2) a discussion of income nonresponse, and 3) a discussion of underreporting of income amounts. One important aspect of the income evaluation not covered in this paper was a reinterview study also conducted by the Bureau. The results of this study were presented at this session in a paper entitled "Problems of Nonsampling Errors in the Survey of Income and Education: Content Analysis," by Robert Fay and Harold Nisselson of the Bureau of the Census. A second paper giving a more general description of the design and field operations of the SIE was also presented at this session in a paper entitled "Description of the Survey of Income and Education Operations," by Marvin Thompson and George Gray of the Bureau of the Census.

CONTRASTING SIE AND THE MARCH 1976 CPS

The Census Bureau has been conducting an annual supplement to the Current Population Survey (CPS) designed to provide annual income statistics for families and persons since 1947. In the spring of 1976, the Census Bureau conducted two surveys yielding estimates of income and poverty for 1975, the annual March CPS and the one-time SIE. Both of these surveys were designed to obtain money income information for calendar year 1975 in a similar fashion.

The results of these two surveys differed significantly in several major areas, the most important difference being in the count of the number of poor and, especially, poor school-age children. Shown in table 1 is a comparison of some selected results of these two surveys. Much effort has been expended to try to explain how and why these differences occurred. Our analysis of these differences started with an enumeration of some of the basic similarities and differences in design of the surveys and in the data collection and processing procedures since these differences helped contribute to the differing results.

SIMILARITIES BETWEEN SIE AND CPS

There are two major areas of similarity in these surveys which, if not similar, would have been prime causes for differing survey results. These are: 1) the design and wording of the labor force, work experience, longest job, and income questions, and 2) the editing and imputation of nonresponses to these questions. The only difference between the SIE and CPS income questions covering calendar year 1975 was that a separate "YES-NO" circle was provided on the SIE for child support payments (it is combined with alimony on the CPS). Since the SIE and CPS questionnaires in the areas of work experience and income were virtually the same, the editing and imputation procedures developed for the March CPS were used to editing and imputation procedures developed for the March CPS were used to edit and impute the SIE.

The decision to use the March CPS income questions on the SIE was made for several reasons. First, given the deadline set by Congress for producing the estimates of poor school-age children by State, the development of a new questionnaire and new processing system, an evaluation based on an independent survey designed to measure the same parameters could be made with at least several important variables held constant. Little or none of the difference between these surveys results from questionnaire wording and design or from editing and imputation procedures.

DIFFERENCES BETWEEN THE SIE AND CPS

Differences between these two surveys which could, and probably did, contribute to some of the differing survey results can be divided into six major areas: 1) survey objectives, 2) sample selection, 3) month of interview, 4) conditioning of respondents, 5) method of interview, and 6) interviewer experience.

SURVEY OBJECTIVES

The stated major objective of the SIE was to collect accurate income information for States with a minimum level of reliability on the estimated number of poor children aged 5 to 17 years, a goal which was for the most part achieved. In contrast, the primary purpose of the CPS is to obtain accurate and timely statistics on the civilian labor force, for example, the Nation's unemployment rate. Collection of income information in the March CPS is acknowledged to be of less importance. This acknowledgement is made during interviewer training but tempered with frequent references to the need for accurate

income data as well.

SAMPLE SELECTION

The method of selection of sample households for these surveys differed considerably. Considerations involving the minimum statistical reliability requirements on the number of poor school-age children by State necessitated a sample design for SIE which differed in several important respects from the March CPS.

The CPS sample is a national multistage, clustered, probability sample made up of self-representing (probability of selection 100 percent) and non-selfrepresenting (probability of selection based on 1970 population) primary sampling units (PSU). PSU's are counties or groups of contiguous counties from which sample households are selected. PSU's from the non-selfrepresenting portion of the CPS sample were chosen from strata formed by grouping PSU's, then selecting, in most cases, one PSU within each stratum to represent that stratum.

The variables used to group non-selfrepresenting PSU's in the CPS into strata included 1) percent urban, 2) percent nonwhite, 3) percent of population employed in manufacturing, 4) SMSA/NON-SMSA, 5) per capita retail trade, 6) rate of population change since the 1960 census, and 7) principle industry. Variables such as per capita income and percent poor were not used.

The SIE sample design was a stratified, multistage, noncompact cluster design. The sample was selected independently within each State. Both selfrepresenting and non-selfrepresenting PSU's were created. Unlike the CPS, the stratifacation of non-selfrepresenting PSU's was largely dependent upon the proportion of poor children 5 to 17 based on the 1970 census. After selection of the sample PSU's, sample ED's within PSU's were selected also with some stratification based on poverty rates from the 1970 census. Finally, within selected ED's, in general, 3 housing units were selected. The poverty status of the sample household and number of children less than 18 years old as of the 1970 census were used to stratify households within ED's before the final sample selection.

MONTH OF INTERVIEW

Traditionally the March CPS is conducted during the week in March containing the 19th. The CPS field collection procedures allow only one week in which to conduct interviews regardless of any supplementary questions such as the March work experience and income questions because of deadlines on release of the monthly national unemployment statistics. A one week followup or extension period beyond this one week using a special income followup form is provided in March in order to obtain information not available during the first week, however, this form is used for only about 5 percent of the interviewed households. In contrast, the SIE interviews took place for the most part in May and June without the one-week time constraint imposed by the March CPS. Since some respondents consult or require tax returns in order to accurately answer survey questions, the SIE would seem to have some advantage over the April 15 filing deadline. The later collection of SIE data may have, however, provided a greater opportunity for telescoping of amounts and presented more serious recall problems for 1975 work experience information and for nontaxable income sources much of which are concentrated in the lower end of the income distribution and therefore important sources of income for the poor.

CONDITIONING OF RESPONDENTS

To assure greater reliability in measuring month to month changes in monthly labor force estimates, the CPS sample consists of eight rotation groups or panels each of which is a national sample. Households in each of these panels are interviewed eight times in two separate 4-month periods in which one interview takes place each month. These two interview periods occur 12 months apart; i.e., a household interviewed for the first time in March 1975 would have been interviewed for the fifth time in March 1976. This overall effect of the conditioning of respondents caused by repeated interviews in the CPS with regard to reporting of income data is not fully known. It is known that the refusal rate, that is, refused to be interviewed rate, increases in the CPS with repeated interviewing. The March 1976 CPS refusal rate of 3.1 percent was, however, somewhat lower than the SIE refusal rate of about 3.5 percent.

MODE OF INTERVIEW

It has been documented that respondent cooperation in answering the income questions in the CPS environment is affected by the method of interview; i.e., personal or telephone interview. Whereas extensive use is made of telephone interviews in the CPS, virtually all SIE data was collected using personal interviews. The lower income nonresponse rate on the SIE (13.0 percent on SIE vs. 19.5 percent on CPS) is probably, to a large extent, related to the exclusive use of personal interviews on the SIE.

INTERVIEWER EXPERIENCE

The large number and wide distribution of the SIE sample households required hiring of a large number of new, temporary interviewers. Most of these "new" hires had no previous experience as interviewers in household surveys. The Census Bureau's permanent staff of interviewers used in the March 1976 CPS was, for the most part, a group of highly trained and experienced personnel who had worked with complex questionnaires and experienced difficult interview situations.

INCOME NONRESPONSE RATES

The level of income nonresponse on the SIE was a major concern to the planners of this survey at the Census Bureau. This concern was especially warranted since the March CPS had been experiencing rapid increases in income nonresponse rates precipitated by the use of inexperienced interviewers used for the SIE, the Census Bureau instituted a very intense quality control operation. This operation was intended to monitor the performance of the interviewers in an effort to quickly correct any problems at the outset before a large number of interviews had taken place.

Since the SIE and March 1976 CPS income questions pertaining to calendar year 1975 were virtually identical and the income data processing system was identical as well, the level of income nonresponse in SIE, would be evaluated by comparison to the March 1976 CPS nonresponse rates. Shown in the first two columns of table 2 is a comparison of the nonresponse rates from SIE and the March 1976 CPS for all persons by type of income (income item).

For purposes of this analysis a person was designated as a nonrespondent if one or more of the 11 income items on the questionnaire for that person were not reported.

The data in table 2 show SIE income nonresponse rates well below the March 1976 CPS nonresponse rates. Overall, the SIE persons nonresponse rate was 1/3 lower than the March 1976 CPS rate. Nonresponse rates were lower for each income type as well. Most of the reduction in the nonresponse rate in the SIE can be attributed to reduction in the number of persons who granted an interview, but refused to respond to all 11 income questions. While the SIE nonresponse rate for persons with one or more, but not all, income responses missing was slightly higher in the SIE (11.8 percent SIE vs. 9.7 percent CPS), only 9 percent of the total number of SIE nonrespondents failed to answer all 11 questions compared to about 50 percent for the March 1976 CPS.

The difference in family income nonresponse rates between SIE and CPS were not so great as for persons (a family was designated as a nonrespondent if one or more family members was a nonrespondent). The March 1976 CPS family income nonresponse rate was 26.0 percent compared to 21.7 for the SIE. This smaller relative difference indicates that income nonresponse in the March 1976 CPS was more concentrated within particular families than in the SIE. About 2.5 percent of all SIE nonrespondent families consisted of all persons failing to answer all income questions. The comparable figure for March 1976 CPS was 45.6 percent.

Although it is difficult to pinpoint and quantify each of the factors which lead to the significantly lower nonresponse rates in the SIE, there are still three factors which may have helped to produce lower nonresponse rates; 1) survey objectives, 2) exclusive use of personal interviews, and 3) the CPS environment.

The first factor, stated survey objectives, is perhaps the most difficult of these factors to analyze. There is no doubt that the SIE had a direct, single, major objective, the collection of income data. The March CPS has several major objectives with the collection of labor force data the most important as stated in the March CPS interviewer training manual. Interviewers in March may jeopardize the labor force statistics if they feel the asking of income questions may result in a noninterview (refusal) when they return in following months. This situation, which can only lead to higher nonresponse rates, did not exist in the SIE.

Based on data from the CPS, the use of personal interviews, as opposed to telephone interviews, results in lower income nonresponse rates. In March 1976 CPS, for example, the persons' income nonresponse rate was 16.1 percent for personal interviews and 21.8 for telephone interviews. About 50 percent of all interviews were personal interviews in the CPS. Aside from the first and fifth month interviews in CPS which are required personal interviews, only about 35 percent of the remaining interviews are personal contacts.

The collection of income data in the CPS environment is probably a third factor indluencing the differences between CPS and SIE income nonresponse rates. Research into the relationship between interviewer experience and income nonresponse on the March CPS has yielded some support to the idea that CPS interviewers trade off higher income nonresponse in an effort to keep low noninterview rates. This study shows some evidence that interviewers with many years of experience administering the March CPS supplement had higher nonresponse rates than interviewers with less experience. CPS interviewer performance ratings are largely based on the interviewer's performance on the current labor force portion of the questionnaire and on the number of refused interviews but is not affected by March CPS income nonresponse rates. One hypothesis is that these interviewers with more experience may be less insistent on obtaining income information if they feel their attempt to obtain income data will result in a refusal when they return the following month.

Some of the interviewing on the SIE was carried out by interviewers who worked on the March 1976 CPS. In all, about 500 of the 2,400 interviewers working on the SIE also worked on the March 1976 CPS. About 20 percent of the interviews were completed by members of this group of CPS interviewers. The nonresponse rates for the group of 500 interviewers with some CPS experience are shown in the right-hand portion of table 2 for the March 1976 CPS and for the SIE.

The data shown in table 2 provide more insight into income nonresponse problems on the CPS than on the SIE. Given the SIE as the collection

vehicle without the constraints discussed earlier involving the CPS environment, the CPS interviewers achieved lower income nonresponse rates on the SIE than on the March 1976 CPS. While in March the CPS interviewers had a 17.1 percent nonresponse rate, they achieved a 13.2 percent rate in SIE.

The SIE income nonresponse rates for persons by State are shown in table 3. Overall higher nonresponse rates were evident in the Northeast and North Central States while the nonresponse rates in the States of the South and West tended to be lower. Although no comparative figures are available for March 1976 CPS by State, data available by Census regional office indicate this same trend.

COMPARISON OF REPORTED INCOME AMOUNTS WITH INDEPENDENT SOURCES

The estimates of aggregate amount of income derived from household surveys are generally deficient (underreported). One method to measure the gross deficiency of income amounts collected in a survey is to compare these amounts to independently derived estimates from administrative sources such as the Internal Revenue Service, Bureau of Economic Analysis, or Department of Health, Education, and Welfare. Figures available from these sources, once adjusted to the Census money income concepts, are valuable in evaluating survey performance. Shown in table 4 is a national comparison of SIE and March 1976 CPS estimates of aggregate income with independently derived aggregate income estimates for 1975 for the income sources covered on the questionnaires of both surveys. A second part of the table disaggregates the "total" survey estimates for each income source into the reported and allocated (imputed amounts due to missing responses) components.

A comparison of SIE and March 1976 CPS estimates of total money income to independent estimates shows the SIE survey yielded an estimate of $1,059.8 billion whereas the March 1976 CPS gave an estimate of $1,017.3 billion. The SIE estimate for sources for which independent estimates are available was 93.7 percent of the independent estimate compared to 90.3 percent for the CPS. This pattern of higher aggregate income amounts in SIE holds for all income sources except alimony and child support.

The significantly lower nonresponse rates in SIE are reflected in the proportion of the total SIE aggregate income which was allocated. While 20.1 percent of the March 1976 CPS aggregate was assigned in the editing and imputation procedures, only 12.1 percent of the total aggregate was assigned in SIE. This difference represents a significant improvement over the CPS.

A comparison of estimated total money income from the SIE and independent sources is shown in table 5 for each State. The independent estimates shown for each State should be considered rough approximations since several figures used to arrive at the national independent estimates were not available on a State by State basis and because the data by State do not reflect recent revisions to the Bureau of Economic Analysis's personal income series. The data indicate that the SIE estimates as a proportion of independent estimates ranged from a low of 87.4 percent in Delaware to a high of 100.7 percent in Arizona.

TABLE 1. SELECTED COMPARISONS OF SIE AND MARCH 1976 CPS INCOME STATISTICS FOR THE UNITED STATES FOR 1975

(Numbers in thousands)

Selected Characteristics	Number			Percent		
	SIE	March 1976 CPS	Difference (SIE-CPS)	SIE	March 1976 CPS	Difference (SIE-CPS)
POVERTY						
Families below the poverty level	5,051	5,450	- 399	9.0	9.7	-0.7
Persons below the poverty level	23,991	25,877	-1,896	11.4	12.3	-0.9
Children 5 to 17 years below the poverty level	7,132	8,034	- 902	14.5	16.3	-1.8
Persons aged 65 and over below the poverty level	3,049	3,317	- 268	14.0	15.3	-1.3
MEDIAN INCOME						
All families	$14,094	$13,719	+$ 375	(X)	(X)	(X)
White	$14,664	$14,268	+$ 396	(X)	(X)	(X)
Black	$ 9,045	$ 8,779	+$ 266	(X)	(X)	(X)
All unrelated individuals	$ 5,168	$ 4,885	+$ 398	(X)	(X)	(X)
All persons with income	$ 5,768	$ 5,664	+$ 104	(X)	(X)	(X)
Men with income	$ 8,974	$ 8,853	+$ 121	(X)	(X)	(X)
Women with income	$ 3,463	$3,385	+$ 78	(X)	(X)	(X)

X Not applicable.

TABLE 2. COMPARISON OF SIE AND MARCH 1976 CPS PERSON'S INCOME NONRESPONSE RATES AND NONRESPONSE RATES FOR CPS INTERVIEWERS WORKING ON SIE, BY TYPE OF INCOME

TYPE OF INCOME	March 1976 CPS	SIE	MARCH CPS INTERVIEWERS WORKING SIE		
			March 1976 CPS		SIE
			Total	MIS 1 [3]	
Total	19.5	13.0	17.1	14.5	13.2
Wages or salary [1]	10.8	6.1	9.3	7.8	6.3
Nonfarm self-employment [1]	7.6	2.5	6.6	5.1	2.5
Farm self-employment [1]	7.2	2.1	6.2	4.5	2.0
Social Security or Railroad Retirement	11.2	2.6	9.6	6.6	2.9
Supplemental Security Income	10.1	1.5	8.5	5.7	1.7
Public assistance or welfare [2]	10.1	1.6	8.6	5.7	1.7
Interest from savings accounts	13.7	7.0	11.9	9.6	7.3
Dividends, rent, estates or trusts	11.7	3.7	10.0	7.4	3.8
Veterans' paymdnts, unemployment compensation, workmens compensation	10.6	2.0	8.9	6.1	2.1
Private, Federal, military, State and local pensions	10.5	1.9	8.9	6.0	2.1
Alimony and child support, contributions from persons not in the household, or any other money income	10.3	1.6	8.7	5.8	1.8

1/ Persons who did not work in 1975 who did not respond to the earnings questions were not considered nonrespondents for these items.

2/ Public assistance and welfare consists mainly of Aid to Families with Dependent Children and General Assistance.

3/ Month-in-sample 1 first interview conducted by personal visit.

TABLE 3. SIE PERSONS INCOME NONRESPONSE RATES BY STATE
(Numbers shown are percents)

State	Nonresponse Rate
Alabama	13.2
Alaska	13.2
Arizona	12.1
Arkansas	9.1
California	12.3
Colorado	12.1
Connecticut	18.3
Delaware	15.2
District of Columbia	10.9
Florida	15.0
Georgia	12.3
Hawaii	9.9
Idaho	12.6
Illinois	16.1
Indiana	12.0
Iowa	12.0
Kansas	13.2
Kentucky	11.9
Louisiana	10.7
Maine	15.4
Maryland	12.3
Massachusetts	15.6
Michigan	13.4
Minnesota	11.6
Mississippi	8.7
Missouri	11.9
Montana	16.0
Nebraska	13.1
Nevada	10.6
New Hampshire	15.0
New Jersey	14.1
New Mexico	9.1
New York	13.4
North Carolina	11.3
North Dakota	13.5
Ohio	15.0
Oklahoma	12.1
Oregon	11.3
Pennsylvania	14.5
Rhode Island	17.1
South Carolina	10.0
South Dakota	12.5
Tennessee	12.0
Texas	11.1
Utah	11.5
Vermont	14.5
Virginia	11.7
Washington	10.7
West Virginia	10.4
Wisconsin	12.8
Wyoming	12.5

TABLE 4. COMPARISON OF SIE AND MARCH 1976 CPS ESTIMATES OF AGGREGATE MONEY INCOME WITH INDEPENDENT ESTIMATES OF AGGREGATE MONEY INCOME ADJUSTED TO CPS MONEY INCOME BY TYPE OF INCOME AND BY REPORTED AND ALLOCATED AMOUNTS

Source of income	Independent sources	March 1976 CPS							SIE						
	Billions of dollars	Billions of dollars	CPS reported and allocated as a percent of CPS total			CPS as a percent of independent sources			Billions of dollars	CPS reported and allocated as a percent of CPS total			CPS as a percent of independent sources		
			Total	Reported	Allocated	Total	Reported	Allocated		Total	Reported	Allocated	Total	Reported	Allocated
Total income	(NA)	1,017.3	100.0	79.9	20.1	(X)	(X)	(X)	1,059.8	100.0	87.9	12.1	(X)	(X)	(X)
Total income, independent estimates available	1,115.6	1,006.9	100.0	79.9	20.1	90.3	72.1	18.1	1,045.2	100.0	87.9	12.1	93.7	82.4	11.3
SOURCES WITH INDEPENDENT ESTIMATES															
Wage or salary income	788.2	767.7	100.0	81.6	18.4	97.4	79.4	18.0	789.9	100.0	89.1	10.9	100.2	89.3	10.9
Nonfarm self-employment income	63.4	61.5	100.0	65.9	34.1	97.0	63.9	33.1	62.6	100.0	77.6	22.4	98.7	76.7	22.1
Farm self-employment	20.9	11.9	100.0	75.6	24.4	56.9	43.1	13.9	13.6	100.0	86.0	14.0	65.1	56.0	9.1
Social Security and Railroad Retirement	65.0	59.1	100.0	79.7	20.5	90.9	72.5	18.6	60.0	100.0	90.8	9.2	92.3	83.8	8.5
Supplemental Security income	5.6	3.6	100.0	86.1	13.9	64.3	55.4	8.9	3.9	100.0	97.4	2.6	69.6	67.9	1.8
Aid to Families with Dependent Children and other public assistance	10.2	7.9	100.0	87.3	12.7	77.5	67.6	9.8	8.1	100.0	95.1	4.9	79.4	75.5	3.9
Interest	59.5	24.7	100.0	70.0	30.0	41.5	29.1	12.4	29.0	100.0	75.5	24.5	48.7	36.8	11.9
Dividends	22.5	11.9	100.0	66.4	33.6	52.9	35.1	17.8	15.3	100.0	70.6	29.4	68.0	48.0	20.0
Net rental income and royalties	11.2	8.2	100.0	76.8	23.2	73.2	56.2	17.0	10.0	100.0	84.0	16.0	89.3	75.0	14.3
Veteran's payments	12.0	8.0	100.0	85.0	15.0	66.7	56.7	10.0	8.6	100.0	93.0	7.0	71.7	66.7	5.0
Unemployment compensation	18.3	11.6	100.0	83.6	16.4	63.4	53.0	10.4	12.4	100.0	94.4	5.6	67.8	63.9	3.8
Workmen's compensation	5.3	2.3	100.0	82.6	17.4	43.4	35.8	7.5	2.4	100.0	95.8	4.2	45.3	43.4	1.9
Private pensions and annuities	13.8	11.2	100.0	75.9	24.1	81.2	61.6	19.6	12.1	100.0	85.1	14.9	87.7	74.6	13.0
Federal government and military retirement	13.5	12.6	100.0	79.4	20.6	93.3	74.1	19.3	12.4	100.0	87.9	12.1	91.9	80.7	11.1
State and local government employee retirement	6.2	4.7	100.0	74.5	25.5	75.8	56.5	19.4	4.9	100.0	91.8	8.2	79.0	72.6	6.5
SOURCES WITH NO INDEPENDENT ESTIMATES															
Estates and trusts	(NA)	2.5	100.0	76.0	24.0	(X)	(X)	(X)	5.3	100.0	73.6	26.4	(X)	(X)	(X)
Alimony and child support	(NA)	3.9	100.0	84.6	15.4	(X)	(X)	(X)	3.0	100.0	93.3	6.7	(X)	(X)	(X)
Contributions from persons not in the household	(NA)	1.7	100.0	82.4	17.6	(X)	(X)	(X)	2.4	100.0	87.5	12.5	(X)	(X)	(X)
All other money income n.e.c.	(NA)	2.3	100.0	78.3	21.7	(X)	(X)	(X)	4.0	100.0	90.0	10.0	(X)	(X)	(X)

NA Not Available.
X Not Applicable.

TABLE 5. COMPARISON OF INDEPENDENT ESTIMATES OF TOTAL AGGREGATE MONEY INCOME FOR 1975 WITH SIE TOTAL AGGREGATE MONEY INCOME BY STATE
(Numbers in millions of dollars)

State	Independent	SIE	$\frac{SIE}{IND}$
Alabama	16,085	14,582	.907
Alaska	2,702	2,636	.976
Arizona	10,749	10,826	1.007
Arkansas	8,921	8,130	.911
California	122,716	117,599	.958
Colorado	13,836	13,380	.967
Connecticut	18,114	17,342	.957
Delaware	3,453	3,018	.874
District of Columbia	4,724	4,281	.906
Florida	41,298	39,738	.962
Georgia	22,285	20,892	.937
Hawaii	5,091	4,691	.881
Idaho	3,860	3,651	.946
Illinois	56,996	49,468	.868
Indiana	26,688	25,926	.971
Iowa	14,563	13,939	.896
Kansas	12,007	11,112	.925
Kentucky	14,449	13,354	.924
Louisiana	16,437	15,332	.933
Maine	4,694	4,379	.933
Maryland	24,214	23,796	.983
Massachusetts	31,854	30,057	.944
Michigan	48,712	46,116	.947
Minnesota	20,500	19,008	.927
Mississippi	8,291	8,220	.991
Missouri	23,161	22,095	.954
Montana	3,674	3,414	.929
Nebraska	8,471	7,457	.880
Nevada	3,592	3,321	.924
New Hampshire	4,139	3,868	.939
New Jersey	44,044	39,847	.905
New Mexico	5,002	4,876	.975
New York	103,439	96,337	.931
North Carolina	23,812	22,738	.955
North Dakota	3,234	2,896	.895
Ohio	55,280	51,861	.938
Oklahoma	12,691	12,294	.972
Oregon	12,052	11,436	.949
Pennsylvania	62,739	56,183	.895
Rhode Island	4,898	4,423	.903
South Carolina	11,490	10,979	.956
South Dakota	3,031	2,729	.900
Tennessee	18,329	17,346	.946
Texas	60,577	56,683	.936
Utah	5,500	5,377	.978
Vermont	2,144	1,994	.930
Virginia	25,877	25,263	.976
Washington	20,442	18,552	.908
West Virginia	8,172	7,537	.922
Wisconsin	23,947	22,519	.956
Wyoming	2,032	1,904	.937

USES OF DATA FROM THE SURVEY OF INCOME AND EDUCATION FOR POLICY ANALYSIS

Alan Ginsburg and George Grob
Office of the Assistant Secretary for Planning and Evaluation, DHEW

Introduction

The Federal government uses data in two ways -- for program administration and for policy analysis. Administrative uses of data are generally sanctioned by law or regulation, and involve the allocation of program resources under established formulas. Such uses are essentially routine and afford little or no room for the exercise of discretion; program entitlements are automatically determined "by the numbers." In contrast, policy analysis, which involves the uses of data to define and evaluate alternative courses of action, is much more episodic and judgmental in character.

Occasionally, when policy analysts focus on existing allocation formulas, the two types of use intersect. Under these conditions, data are used analytically to evaluate other administrative uses of data. Such has been the case with the Survey of Income and Education (SIE). The survey was expressly authorized by Congress with a view to evaluating the continued use of 1970 Census poverty statistics in the allocation of funds under Title I of the Elementary and Secondary Education Act.1/

The question of the likely impact of the SIE data on Title I allocations will be examined in the second half of this paper. Here, we wish to consider briefly the variety of Federal programs already using data similar to that now available from the SIE, and to describe some of the specific ways in which the SIE data lend themselves to policy studies.

Administrative Uses

Overall, programs in at least five departments -- Agriculture, HEW, HUD, Labor, and Treasury -- utilize income or employment statistics in current allocation formulas, and thus, are potential users of the SIE data. In terms of total funding, the Revenue Sharing program administered by the Treasury Department is the largest. Funds are allocated to States and local governments based on interdependent formulas in which the key variables are population, per capita income, and adjusted taxes.

In the areas of employment, there are two major programs -- the Comprehensive Employment and Training Act administered by Labor, and the vocational training program operated by the Office of Education in HEW. Both programs define eligibility and apportion assistance on the basis of poverty measures and unemployment rates. Other programs utilizing poverty measures as a basis for distributing aid are the Department of Agriculture's Food Stamp Program, the Community Mental Health Centers program of the National Institute of Mental Health, and HUD's Title I program under the Housing and Community Development Act of 1974. The latter program is designed primarily to improve housing for low to moderate income families in metropolitan areas.

All of these programs have one problem in common: significant changes have occurred since the 1970 Census, and as a result, there may be serious inequities in current patterns of assistance. Some of these changes, such as regional and metropolitan migration trends, are already well-documented, thanks to the Current Population Survey, but now the SIE offers reliable estimates of the net effect of these trends on the demographic and economic characteristics of individual States.

Will the SIE data supplant 1970 Census figures in existing allocation formulas? In some cases, such as Revenue Sharing, the answer is clearly no, since the SIE cannot begin to provide adequate estimates for the thousands of local governments involved. In other cases, programs may be wedded to 1970 poverty estimates because of required linkages with other types of data available only from the Decennial Census. This may apply to urban redevelopment programs, for example, since the SIE provides almost no information on housing characteristics.

As a general proposition, program managers are reluctant to make any change in grant procedures without a thorough study of the consequences, both statistical and political. With major tabulations of the SIE results now available, the statistical consequences are largely known, but more time may be required to assess the political ramifications. In this connection, it must be observed that delay serves to bring closer the time when the 1980 Census will make the question of using the SIE data entirely moot. Finally, two major evaluation studies bearing on the reliability of the SIE data and methods of developing sub-state poverty estimates are just now being completed.2/ These studies were mandated at the same time as the Survey, to guide the Congress in its deliberations on updating the Title I allocations. Clearly, the precedent set in this program area will carry considerable weight throughout the government.

Contributions to Policy Analysis

Before turning to an examination of some of the issues that are likely to shape the Title I decision, we wish to offer a few observations concerning the exceptional value of the SIE data for policy analysis.

Perhaps the first thing to be said is that the value of the SIE data to program planners and administrators is no accident: they played a major role in specifying the survey content. As a result, questions were added dealing with Food Stamp and public assistance recipiency, housing costs, liquid assets, child and adult disability, public and private health insurance coverage, and

changes in family composition affecting income reported for the previous year.

Broadly speaking, three types of policy analyses are being carried out: methodological studies, diagnostic studies, and simulation studies. Methodological studies have focused primarily on alternative measures of poverty. One of the richest areas in the SIE data for diagnostic studies is that of labor force participation, since extensive probes are utilized to develop a comprehensive picture of employment, job seeking, and reasons for periods of non-employment. Simulation studies utilizing SIE data have been devoted largely to evaluating welfare reform proposals. Working with data for individual households, and drawing on known relationships among various socio-economic variables such as age, occupation, and income, it is possible to simulate and hence "cost-out" or quantify the effects of different eligibility criteria and benefit levels. The ability to treat State of residence as a variable in these simulations has greatly improved their accuracy, since most existing welfare programs, including Aid for Dependent Children, Medicaid, and Food Stamps, are administered by the States, and benefit levels vary widely.

The SIE will continue for some years to be a prime source of data for policy analyses, but it is still reassuring to know that a successor survey is already in the works. Based on recommendations stemming from a comprehensive review of income statistics conducted by the Office of Management and Budget in the Spring of 1973, the Department of Health, Education, and Welfare is now planning a recurring Survey of Income and Program Participation, in conjunction with the Census Bureau. Many features of the SIE will be incorporated into the new survey, since the data, interview forms, sampling techniques, and field experience from the SIE have been consulted extensively in its development. The survey is slated to become operational in the 1980's.

Title I, ESEA

Although SIE data bear on policy issues in a number of different Federal programs, the survey's Congressional mandate was solely to update the poverty criterion specified in Title I of the Elementary and Secondary Education Act. Based on difficulties experienced earlier in converting from 1960 to 1970 Census estimates of poor families, the Congress decided that ten years is too long an interval between updates.

Title I participants are selected within school districts on the basis of various measures of educational need, but the estimated number of children in low income or "poverty" families is the key variable in the allocation of funds to States and local areas.3/ Thus, with the results of the SIE now in, it is possible to "cost out" the implications of changes since 1970 in the distribution of poverty children for Title I payments to the States. In this connection, it should be noted that Congress reserved for future deliberations the question of whether SIE estimates should supplant the 1970 figures in the Title I formula. These deliberations have now begun, and it is already clear that evaluations of the SIE results will figure conspicuously in the debate.

The SIE results show that there has been a significant shift in the distribution of children in poverty families between 1970 and 1976, based on income for the years 1969 and 1975 (Table 1). Comparatively fewer children now reside in the South (minus 12 percent), with the largest decreases occuring in Alabama (46 percent), Arkansas (26 percent), Louisiana (23 percent), and West Virginia (23 percent). Comparatively more poverty children reside in the industrial States. States showing the largest increases are New Jersey (35 percent), Illinois (39 percent), Michigan (23 percent), Pennsylvania (16 percent), and Ohio (16 percent). Among the twelve smallest States, seven show changes in excess of 30 percent, with large increases observed in Nevada, New Hampshire, and Vermont, and large decreases observed in Alaska, the District of Columbia, and the two Dakotas.

If sanctioned by Congress, these changes would have roughly proportional effects on Title I allotments. Based on FY 1977 figures, $131 million would have been re-allocated among States. While this is only about 8 percent of the total funding, the impact on individual States is considerable. In the two extreme cases Illinois would gain $27 million and Alabama would lose $22 million. Alabama would also experience the greatest proportional decrease (48 percent), while Nevada would receive the largest relative increase (54 percent).

Given the magnitude of these potential impacts, political considerations are likely to outweigh statistical ones in the final decision of the Congress. Nevertheless, statistical evaluations of the SIE results will figure in the debate, and in this connection there are three issues which are likely to receive close attention. These involve questions of sampling error, income reporting, and use of the SIE data for county-level estimates.

Despite the large size of the SIE sample -- over 150,000 households were interviewed -- the possibility of errors associated with sampling are likely to weigh heavily with Congress, particularly when translated into Title I allotment amounts. Since the sample was designed to minimize the relative error of the State estimates (the objective was to keep the coefficient of variation under 10 percent), the size of the absolute error in the larger States can be considerable. In the case of California, for example, one standard error in the estimate of children in poverty translates into $9.8 million of Title I funds, based on FY 1977 allotments. This amounts to more than two-thirds of the total cost of the survey. Thus, some will argue that there is a serious disproportion between the accuracy of the SIE (and its associated costs) and the amounts at risk under the Title I program.4/ Given the logic of the SIE sample design, there are a great many statements which can be made concerning the likelihood of error, some of which will doubtless excite concern in the Congress. Thus, for ex-

TABLE 1: CHANGES IN RELATIVE SHARES OF POVERTY CHILDREN AND TITLE I FUNDS FOR STATES AND REGIONS: 1970 TO 1976

REGION State	Percent of total poverty children in United States			Title I allocations for FY 1977 (in millions)		Difference	
	1970 Census	SIE (1976)	Percent increase	Actual (based on '70 Census)	Hypothetical (using SIE)	Amount	Percent
United States, Total	100.00	100.00	--	$ 1,653	$ 1,653	--	--
NORTHEAST	16.21	18.64	15.0	380	417	37	+10
Maine	0.47	0.54	14.9	6	7	1	+16
New Hampshire	0.19	0.28	47.4	3	4	1	+32
Vermont	0.17	0.28	64.7	3	5	2	+48
Massachusetts	1.52	1.73	13.8	32	35	3	+9
Rhode Island	0.32	0.30	- 6.3	7	6	-1	-12
Connecticut	0.72	0.83	15.3	15	18	3	+20
New York	6.84	7.35	7.5	184	189	5	+3
New Jersey	2.02	2.72	34.7	46	60	14	+32
Pennsylvania	3.96	4.61	16.4	84	93	9	+11
NORTH CENTRAL	20.00	22.12	10.6	363	400	37	+10
Ohio	3.55	4.12	16.1	52	58	6	+11
Indiana	1.60	1.69	5.6	21	23	2	+9
Illinois	3.93	5.46	38.9	88	115	27	+31
Michigan	2.86	3.53	23.4	70	84	14	+19
Wisconsin	1.35	1.49	10.4	27	29	2	+6
Minnesota	1.28	1.22	- 4.7	25	25	--	-2
Iowa	0.94	0.75	-20.2	15	12	-3	-20
Missouri	2.25	2.20	- 2.2	31	29	-2	-4
North Dakota	0.36	0.25	-30.6	5	3	-2	-34
South Dakota	0.44	0.30	-31.8	5	4	-1	-24
Nebraska	0.60	0.51	-15.0	10	8	-2	-24
Kansas	0.84	0.60	-28.6	14	10	-4	-34
SOUTH	49.55	43.43	-12.4	660	571	-89	-13
Delaware	0.23	0.20	-13.0	5	4	-1	-18
Maryland	1.52	1.46	- 3.9	28	28	--	--
District of Columbia	0.48	0.32	-33.3	10	7	-3	-31
Virginia	2.78	2.18	-21.6	39	31	-8	-21
West Virginia	1.38	1.07	-22.5	18	12	-5	-27
North Carolina	4.06	3.10	-23.6	50	39	-11	-22
South Carolina	2.69	2.32	-13.8	34	29	-6	-17
Georgia	3.82	3.57	- 6.5	50	44	-6	-11
Florida	3.89	5.36	37.8	61	80	19	+31
Kentucky	2.71	2.39	-14.8	34	30	-4	-12
Tennessee	3.18	2.74	-13.8	41	34	-7	-16
Alabama	3.53	1.91	-45.9	46	24	-22	-48
Mississippi	3.40	2.72	-20.0	42	34	-8	-20
Arkansas	2.01	1.49	-25.9	26	18	-8	-29
Louisiana	4.01	3.08	-23.3	53	39	-14	-27
Oklahoma	1.59	1.22	-23.3	19	15	-4	-21
Texas	8.27	8.30	0.3	104	103	-1	-2
WEST	14.26	15.80	10.8	248	265	17	+7
Montana	0.32	0.32	0.0	5	5	--	-4
Idaho	0.31	0.32	3.2	4	4	--	+4
Wyoming	0.13	0.11	-15.4	2	2	--	-4
Colorado	0.93	0.90	- 3.2	16	14	-2	-11
New Mexico	1.05	1.09	3.8	14	14	--	-4
Arizona	1.09	1.30	19.3	15	18	3	+17
Utah	0.40	0.35	-12.5	6	5	-1	-13
Nevada	0.14	0.22	57.1	2	3	1	+54
Washington	1.04	1.14	9.6	19	21	2	+11
Oregon	0.70	0.60	-14.3	15	13	-2	-17
California	7.74	9.09	17.4	141	158	17	+12
Alaska	0.16	0.09	-43.7	3	2	-1	-23
Hawaii	0.25	0.27	8.0	6	6	--	--

ample, the probability of ten or more State estimates being off by more than 10 percent is .99, and conversely, the chance of estimates for all 50 States and the District of Columbia being within 10 percent is one in a billion.

It is interesting to note that while the SIE sample design equitably distributes the risk of error across States, it is not efficient from the standpoint of targeting Title I dollars. To minimize the number of Title I dollars misdirected as a result of sampling error, the sample size would have to have been proportional to the estimated number of children in poverty as well as the inverse of the poverty rate.5/

Based on reinterview studies, it appears that income reporting in the SIE was relatively more complete than that generally obtained in the Decennial Census or the Current Population Survey. Since the effect of unreported income is to inflate estimates of poverty, this means that the SIE provides the best estimate of the total number of children age 5-17 in families falling below the poverty line, but it also means that SIE-Census comparisons provide a distorted picture of changes over time. Using the results of the March CPS for 1970 and 1976 as a bridge between the 1970 Census and the SIE, we estimate that if the rate of nonreporting of income in the SIE had been comparable to that in the 1970 Census, the number of children in poverty would have been 23 percent higher. This means that instead of the observed decrease of 7.4 percent since 1970, we might have obtained an increase of 14.3 percent (Table 2), a shift of nearly 22 percentage points. While these calculations are admittedly speculative, and depend on the assumption that no change in income reporting has occurred between 1970 and 1976 in the CPS, the potential magnitude of these shifts is enough to justify serious concern. The basic intent of the SIE was to obtain estimates which would help to bridge the gap between the 1970 and 1980 Censuses, but on this evidence, it seems likely that the 1980 Census will show substantially greater numbers of children in poverty than could have been expected from the SIE results.

One final point should be noted concerning the question of sampling error. Even when the 1970 Census figures for children in poverty are rateably reduced to yield the same total as the SIE, most of the major changes observed at the State level are highly significant. Thus, for Illinois and New Jersey -- the two States which stand to gain the largest amounts -- the observed differences are respectively 4.0 and 3.3 times the standard error of the estimates. This also holds true for States which would experience the largest relative increases: Vermont, Nevada, and New Hampshire all exhibit differences in excess of 3.5 standard errors.

As we have indicated, there is some question of whether the SIE estimates are sufficiently accurate at the State level, and the Title I allocation process requires estimates of poverty children down to the county level. Thus, if Congress were to authorize the use of the SIE estimates in determining the amount each State receives, the problem of how the States would suballocate to the county level still remains. One possible solution, explored by Abduhl Kahn and Herman Miller in connection with a special study mandated by Congress, is to develop synthetic estimates. This method applies trend data for metropolitan and non-metropolitan counties at the State level as an adjustment to 1970 data for individual counties. Two serious limitations of this approach are: 1) measures of reliability are not calculable for such estimates, and 2) the method may be perceived as open to manipulations designed to produce preconceived results. Logically, the easy way out would be to let the States work out their own methods of allocating funds down to the county level, but Congress is very reluctant to do this, for fear that some States would re-direct Federal funds away from low-income areas.

TABLE 2: ADJUSTMENTS IN SIE ESTIMATE OF POVERTY CHILDREN FOR COMPARABILITY WITH 1970 CENSUS, BASED ON COMPARISONS WITH CURRENT POPULATION SURVEY ESTIMATES.

	Related children 5-17 in poverty (thousands)		
	Actual estimate	Adjusted estimate*	Percent change
1970 Census.....	7,700	7,700	--
1970 March CPS....	7,000	7,700	+10.0
1976 March CPS....	8,000	8,800	+10.0
1976 SIE.....	7,132	8,800	+23.4
Percent change '70 Census to SIE.....	-7.4	+14.3	

* Estimates are adjusted for comparability with the 1970 Census. CPS income reporting is assumed to have remained the same between 1970 and 1976; thus, our calculations suggest that the 1970 Census methodology would have yielded an estimate 10 percent higher than the CPS in 1976 just as it did in 1970.

Conclusion

Data from the SIE will be used over the next four to five years by policy analysts in a number of different program areas within the Department of Health, Education, and Welfare, including bilingual education, education of handicapped children, welfare reform, and postsecondary education. As mentioned earlier, at least four other departments -- Labor, Treasury, Agriculture, and Housing and Urban Development -- also plan to make special use of the SIE data. Clearly, then, there is no question about the benefits of the survey exceeding its cost. Relative to the annual appropriations of the programs benefiting, the $14 million cost of the SIE is a nearly invisible fraction. Relative to the original purpose of updating the Title I allocations, however, these are "fringe" benefits, based largely on add-ons to the scope of the survey.6/ Thus, while the

Federal government can congratulate itself on having successfully exploited the opportunity afforded by this mandated survey, the possibility that it may never be used for the purpose originally intended must give pause for reflection.

In retrospect, it was a mistake for the Congress to defer a decision on the use of the SIE. As a result, the question of the required degree of accuracy of the SIE estimates was not conclusively resolved, and now the need for further deliberations means that Title I allocations cannot be updated until FY 1980 at the earliest -- just two years from the time when the results of the 1980 Census will become available.

At the present time, it appears that the SIE data will be used to up-date State allocations, but reaching agreement on this is likely to require a hold-harmless provision plus an increase in Title I funding. Based on the funding level in FY 1977, a full hold-harmless would cost an additional $131 million on a base of $1.6 billion. Within-State allocations will probably continue to be determined on the basis of the 1970 data for counties.

Looking to the future, there is a serious question of whether the 1980 Census will produce poverty estimates comparable to those of the SIE. If past experience is any guide, under-reporting of income is likely to inflate the Census estimates of poverty. There is even some question as to whether the 1980 income data will be comparable to those from 1970, since the Census Bureau is experimenting with a simplified income question for use on the complete-count form, and plans to ask much more detailed income questions in a follow-on survey.

With the authorization of a mid-decade or quinquennial census, major Federal programs will no longer have to endure a ten-year hiatus between reliable measures of the social and economic conditions they are designed to ameliorate. It is likely, however, that special surveys will continue to be required in order to provide the necessary detail for narrow-gauge programs targeted on particular types of disadvantaged groups. In this connection, we believe the SIE will serve as a useful model of the benefits to be realized by pooling needs and sharing costs.

FOOTNOTES

1/ Section 822a of the Education Amendments of 1974 -- Public Law 93-380.

2/ These reports, now being prepared at HEW and the Census Bureau, are entitled (1) "The Survey of Income and Education" and (2) "Counting Poor School Children".

3/ Slightly simplified, the allocation formula may be described as "eligibles" times "payment rate" times "rateable reduction", where: (1) eligibles are the sum of the children age 5-17 in poverty families as defined in the 1970 Census plus two-thirds of children in families above the poverty line line receiving AFDC payments plus children in foster homes or institutions for the neglected and delinquent; (2) the payment rate is 40 percent of each State's current educational expenditure per pupil but not less than 80 and not more than 120 percent of national average expenditures, and (3) the rateable reduction is the ratio of the current appropriation to the amounts otherwise authorized.

4/ Achieving a coefficient of variation of 2½ percent was estimated to cost between $50 and $100 million. This was judged to be excessive, in part because added costs would have come out of Title I program money, and the program is already funded at substantially below the estimated level of need.

5/ To minimize the variance of the individual State estimates, the fraction of the total sample (n) allocated to a given State (s) with C_s estimated children in poverty constituting Ps proportion of all children is given by the proportion:

$$\frac{C_s}{C_n} \sqrt{\frac{P_s \quad (1-P_s)}{P_n \quad (1-P_n)}}$$

6/ In the case of data needed for the bilingual education program, it was necessary not only to add questions dealing with limitations in the use of English, but also to expand the SIE sample in selected States in order to obtain estimates of sufficient reliability for children of limited English-speaking ability.

PROBLEMS OF NONSAMPLING ERROR IN THE SURVEY OF INCOME AND EDUCATION: COVERAGE EVALUATION

Eli S. Marks and Harold Nisselson, U. S. Bureau of the Census

The primary purpose of the Survey of Income and Education (SIE) was to estimate by state, the number of children aged 5-17 in poverty families. Since the SIE was conceived as an exploration of the feasibility of using intercensal estimates of children in poverty families from a sample survey for allocating Federal aid to education funds, the legislation authorizing the study also contains a requirement that there be an evaluation of the accuracy and utility of the SIE results. Part of this evaluation involves the estimation of nonsampling error effects on the survey estimates.

Apart from sampling error, estimates of the number of children in poverty families are affected by "content errors" in reporting income and age and by "coverage errors" (primarily omissions) in reporting persons and housing units. To study content errors, a subsample of the housing units included in the SIE was selected and reinterviewed. This content evaluation by use of a Reinterview Sample is discussed in another paper presented at this session.[1/]

The SIE Reinterview Sample was also used to estimate coverage error due to the omission of persons in housing units included in the original SIE sample. This also included a check on the coverage of persons in SIE sample households that were erroneously classified as vacant. There were, however, coverage errors due to omission of housing units from the SIE sampling frame. To check on the coverage of housing units (i.e., on housing units omitted from the sampling frame), a coverage check was carried out on a sample of housing units linked to the SIE Reinterview Sample.

Obviously, a sample of housing units to check on the sample frame coverage has to include housing units not in the original sample frame. Ordinarily this involves selecting a sample of areas (segments), listing all the housing units in the sample areas and determining which of the listed housing units are in the sampling frame, i.e., had a chance of being included in the sample. However, the costs and problems of delineating sample areas of satisfactory size for a housing unit coverage check were substantial. It was, therefore, decided (a) to use alternative sampling procedures which did not require delineating (exact) boundaries for sample areas; and (b) to restrict the coverage check to those sections of the population where we would anticipate substantial undercoverage. It was also decided not to check the coverage in the 'New Construction' and 'Special Places' strata of the sampling frame since, for these strata, the difficulties and costs of matching a listed housing unit to the frame would be very considerable and the yield in terms of missed housing units was expected to be small.

The Within Structure Listing Check

For purposes of the SIE coverage check, two classes of missed housing units were defined--(1) missed housing units in enumerated structures (i.e., in structures included in the SIE sampling frame) and (2) housing units in missed structures. Missed housing units in enumerated structures obviously involve multi-unit structures or structures which existed in 1970 and had residential quarters but which have been 'converted' to some other housing unit layout since 1970.[2/] Thus, they involve housing units which existed in 1970 but were missed by the 1970 Census and housing units (or non-housing unit living quarters) created since the 1970 census in structures built prior to the census. The missed units in converted structures are mostly in urban areas (particularly in central cities of SMSA's). The other missed units in enumerated structures are also mostly in urban areas since they involve multi-unit structures. Special problems exist in measuring coverage errors associated with converted enumerated structures since conversion can reduce, increase or leave unchanged the number of housing units in a structure.

To check on missed units in enumerated structures, the structures in which each of the Reinterview Sample (regular) housing units were located were relisted and the relistings for any multi-unit structures (shown as "multi-unit" either in the 1970 Census Address Registers or in the relistings) were matched to the Address Registers and the missed units were identified. Since the original SIE sampling procedure provided for relisting and resampling multi-unit structures where the sampling unit originally selected could not be identified,[3/] structures which had been relisted for the original SIE were omitted from the within structure coverage check. However, this left a substantial number of multi-unit structures in which there was trouble in identifying each of the housing units listed in the Census Address Register with a corresponding unit on the Within Structure Listing form. All such structures were treated as 'converted structures'.[4/] A sample unit (or units) was selected for interview within the 'converted' structure, the effect of "net coverage error" being defined as the difference between the results obtained in the coverage check interview(s) and the results obtained in the original SIE interview.[5/]

Where all the Address Register listings for a multi-unit structure matched units on the Within Structure Listing (WSL) form but there were additional units on the WSL form, these additional units were identified as 'missed housing units' and interviews taken to determine the characteristics of the occupants.

Where a structure that contained an SIE Reinterview Sample unit had more than 12 housing units, it was to be subdivided (by the SIE reinterviewer)

into 'chunks' (floors, wings, etc.) with 12 or less housing units and only one of these 'chunks' (the one containing the original sample unit) was to be listed. Thus, for larger structures, the relisting was of a subsample rather than the entire structure. This does not alter the basic procedure for estimating 'within-structure misses' but merely the specific sampling probabilities involved in making the estimate.

The Successor Structure Check

The SIE sampling procedure departed from that used in the CPS with respect to:

1) Sample selection in multi-unit structures:- Here, CPS relists and resamples multi-unit structures whenever a unit from such a structure is selected for the sample. The SIE relisted and resampled multi-unit structures only when it was not possible to identify the housing unit originally selected for the sample.

2) In rural areas and other areas without clearly identified addresses (street or road names and house numbers), the CPS selects a sample of small areas (segments). The SIE selected individual housing units from the 1970 Census Address Registers (just as in the "address E.D.'s") and the interviewers located these on the basis of whatever information was available (name of 1970 household head, box number, E.D. map spotting).

The Within Structure Listing already described gives an estimate of the effects of the relatively minor modification in the procedure for sampling multi-unit structures. More important from the standpoint of future sampling methodology at the Census Bureau, was a check on the effects of the change in rural areas and small towns from a segment (area) sample to a list sample. To provide a measure of these effects, a Successor (Structure) Check was done. Here, the SIE reinterviewer was instructed to start from the structure containing the reinterview sample unit,[6] (where he reinterviewed the sample unit and completed a Within Structure Listing form); and, proceeding always to the right without crossing a street or road unless it came to a dead end, to list all the structures he encountered until he had listed, in addition to the sample unit, four 'successor' structures built before the 1970 Census. The SIE reinterviewer was to list the names of the current household head(s) and the head(s) in 1970 and any address or description and to check whether the unit was built before or after the 1970 Census.

As shown in Table 1, the 1970 Census housing coverage check[7] had indicated that omissions of entire structures from the Census was most common in rural areas and in urban E.D.'s outside urbanized areas. It also shows a relatively small missed rate for housing units in missed addresses in the larger urban places.[8] For this reason, the Successor Check was restricted to rural areas and to urban places of less than 10,000 population outside of urbanized areas.

In the rural areas and the very small urban areas, 'addresses' are usually not specific to a structure and are, therefore, not of much use for matching. Determination of which successor structures were listed in the 1970 Census Address Registers had to depend on matching names of 1970 household head and map locations of the structures. The map locations were obtained for the Census from the instruction given in 1970 to enumerators in rural areas to draw a small box on the E.D. map to indicate the location of each structure listed in the Address Register, labeling it with the Census Serial Number(s) for the structure. For the Successor Check, the reinterviewers were instructed to draw a sketch map, labeling roads, streams, etc., and 'spotting' each structure listed by them on this sketch map.

Table 1
Estimated Missed Housing Units per 100 Enumerated Units, 1970 Census of Housing and Population

	Total Missed Units	Missed Units In Enumerated Addresses	Missed Units in Missed Addresses
Total U.S.	2.5	0.5	2.0
Rural	4.8	0.2	4.6
Urban	1.7	0.6	1.1
In Urbanized Area	1.3	0.5	0.8
Outside Urbanized Area	3.1	0.8	2.3
Places by Size			
1,000,000 and over	1.1	0.8	0.3
500,000-999,999	0.1	N.A.	0.1
250,000-499,999	0.9	0.4	0.5
100,000-249,999	2.3	1.8	0.6
50,000-99,999	1.7	0.4	1.3
25,000-49,999	2.0	0.6	1.4
10,000-24,999	1.5	0.7	0.8
2,500-9,999	3.2	0.5	2.6

Note: Above are field enumeration coverage rates only (before corrections made in processing). They are taken from The Coverage of Housing in the 1970 Census, U.S. Bureau of the Census, Census of Population and Housing; 1970, Evaluation and Research Program PHC(E)-5. Sampling errors and descriptions of methodology and limitations of the 1970 housing coverage studies appear in the report PHC(E)-5.

In many cases, matching was impossible because names of 1970 household heads were missing or incorrect on the Successor Check listings and map spottings (particularly for the 1970 Census listings) were absent, inaccurate or illegible.[9] It was necessary, therefore, to send over a third of the SC forms back to the field for 'reconciliation'. In the reconciliation, the interviewer was told to try to obtain more definitive information (primarily names of all possibilities as 1970 household heads for the unit) to determine whether the structure was or was not listed in the 1970 Census Address Register and to continue the SC listings until a total of four successor structures which appeared on the 1970 Census Address Registers had been listed (or until certain cut-offs, established for the original SC listing, had been reached). To avoid having to send back for a second or third 'reconciliation' cases not resolved by the first 'reconciliation', the

reconcilers were given copies of the 1970 Address Register Sheets which contained the sample unit and the structures near it. They were also told to get interviews for the housing units they determined to be missed if there were one or two such units. Interviews were not taken for cases with 3 or more unmatched units because subsequent matching in the office usually indicated that such cases were matched by housing units listed on Census Address Register sheets not supplied to the reconciler.

Discussion

As already noted, a coverage evaluation was considered important for the SIE from both the methodological and substantive standpoints. From the methodological standpoint, the SIE introduced some changes over the CPS sampling procedure and it was, therefore, desirable to check whether the coverage resulting from these changes was satisfactory. From the substantive standpoint, the undercoverage could have an important impact on the count of children in poverty, and its distribution among states (and between urban and rural areas) because of the greater missed rates usually found for low income families. It was, in fact, possible that the content and within household coverage checks carried through on the Reinterview Sample proper would tend to reduce the counts of poverty families and of children in such families. That is, more family incomes will tend to be adjusted upward than downward due to (a) the reporting of previously omitted income sources and (b) due to adding omitted income recipients. It is true that the within housing unit coverage check would tend to increase family size but a large component of within household undercoverage is the omission of adult male wage earners.[10] As opposed to this, children ages 5 to 14 tend to be exceptionally well-enumerated among blacks and, probably, among most other groups. However, household reinterviews are usually unsuccessful in detecting missed adult males in enumerated low income families and there is no assurance that the reinterviews with enumerated households in the SIE sample will adequately measure the effects of content errors and within household coverage error.

In contrast with the upward bias of the estimates of the number of children in poverty families due to errors in income reporting and within household coverage, we would expect a downward bias due to the omission of housing units and, because of the greater omission rates for lower income households, we would also expect downward bias in the estimates of the proportion of all children who are in poverty families.

With respect to the estimated coverage of housing units in enumerated structures, the SIE compares favorably with the Census (and, probably, with the CPS). The estimated rate of missed housing units (per 100 enumerated housing units) in enumerated structures is 0.5%, which is the same as the coverage rate of missed housing in enumerated structures estimated for the U.S. as a whole over the 1970 Census. This may, in fact, represent an improvement of the SIE in the 1970 Census coverage within enumerated structures, since the Within Structure Listing coverage estimate should include some housing units "converted" to residential use after the 1970 Census as well as housing units actually missed by the 1970 Census. It is likely that any difference of this type is due more to sampling and matching error than to improved SIE marksmanship.

The picture for the change from an area to a list sample in rural areas and small towns, is less encouraging. For these areas, preliminary SC coverage check estimates are of the order of 7 to 13 missed housing units in missed structures per 100 enumerated housing units (in enumerated structures).[11] This is greater than the 4.6 missed units per 100 enumerated units in missed structures reported for rural areas in the 1970 Census and the rate of 2.6 missed housing units per 100 reported for urban places of population size 2500 to 9999. There is an excellent chance that the SC missed rate represents a real difference in coverage between a "segment" sample and an address sample (of individual housing units built before 1970) due to 'conversions' and particularly 'conversions' of vacant structures which were considered to be nonresidential or "unfit for human habitation" in 1970 because at the time they were vacant but which were classified as housing units because they were occupied for residential use at the time of the SIE.

If, as is likely, a missed housing rate of 13 per 100 (or even of 7 per 100) is considered unsatisfactory for Bureau of the Census surveys and this makes a straight list sample of (enumerated) addresses infeasible, we may want to consider a successor sample as an alternative to a regular area sample, provided we can solve the cost problems associated with making additional visits to 'reconcile' matching problems and to 'complete the string' of units listed in the original sampling frame. That is, the very marked increase in recent years in the costs of delineating satisfactory area segments for sampling purposes may more than offset the 'successor sample' costs of doing a moderate amount of revisits for 'reconciliation' and 'completing the string' of enumerated structures.

It should be noted that the successor check used in the SIE represents a modification of the 'half-open interval' approach used for some previous coverage checks. In the 'half-open interval' approach, units (if any) from the starting point through the next previously listed unit are in the sample. This was modified for SIE to extend the sample 'segment' through the next four previously listed units since, while extending the listing means increased cost, it also means a more than proportionate reduction in variance. The Census Bureau is planning to analyze the data from the SIE successor check and other successor checks done subsequently, to try to estimate the optimum cluster size (from the cost-variance standpoint).

While the successor checks used to date have been used for checking on the coverage of a housing unit listing, the technique can, of course, be

used for updating an old listing. The procedure could be used for the purpose of list updating without matching to the old list by determining those structures which should have been on the old list. This involves carrying forward to the new listing the undercoverage of the old one. In theory, the procedure is, in other respects, no more biased than the listing of an area segment, since the errors made by successor listers in following the route and defining old and new construction, correspond to the errors made by area listers in defining the area boundaries and covering all the units inside those boundaries and none outside of it.

Footnotes

1/ Problems of Nonsampling Error in the Survey of Income and Education: Content Analysis by Robert E. Fay III.

2/ These conversions are not included in the 'New Construction' strata.

3/ This happened either because of inadequate distinction in the Address Register between the housing units at the address or because of conversions or changes in the housing unit identification system.

4/ Many of these cases are merely failure of the housing unit designations to correspond--e.g., one listing shows 1st floor right, 1st floor left, 2nd floor right, 2nd floor left and the other shows apartments 1, 2, 3, 4.

5/ Where the original SIE sample unit was selected for a coverage check interview, the "net coverage error" was defined as zero and no coverage error interview was taken.

6/ Where the sample unit is in a multi-unit structure, one must also allow for the probability that the structure (or structure 'chunk') will be in the sample. This probability is, of course, proportional to the number of units listed for the structure (or structure 'chunk') in the 1970 Census Address Register.

7/ U.S. Bureau of the Census, Census of Population and Housing: 1970, Evaluation and Research Program PHC(E)-5, The Coverage of Housing in the 1970 Census, U.S. Government Printing Office, Washington, D.C., 1973

8/ As might be anticipated, rural areas and the smaller urban places show low rates for housing units missed at (multi-unit) enumerated addresses.

9/ Many of the Census E.D. maps were of such small scale that it was impossible to distinguish between the locations of individual houses in a row of 5 to 10 successive structures.

10/ This shows up clearly in the much higher undercounts in most U.S. censuses and surveys for black males than for black females in the age range 20 to 54.

11/ The range reflects the serious difficulties (and the resultant uncertainties) encountered in trying to match housing units in areas where information on address or location is missing, vague or erroneous. The figures cited are subject to sampling error.

PROBLEMS OF NONSAMPLING ERROR IN THE SURVEY OF INCOME AND EDUCATION: CONTENT ANALYSIS

Robert E. Fay III, U.S. Bureau of the Census

Introduction

Congress mandated the 1976 Survey of Income and Education (SIE) through the legislative injunction to "expand the current population survey (or make such other survey)" to furnish current estimates by State of the number of school age children living in poverty families. The legislation directed that these estimates be analyzed for possible use in the allocation of educational monies to school districts under Title I of the Elementary and Secondary Education Act of 1965. The Congress further enjoined the Secretaries of Commerce and HEW to submit a report on the survey, "including analysis of its accuracy and the potential utility of the data derived therefrom..." for updating this allocation. By agreement between the two departments, the Bureau of the Census assumed responsibility for the analysis of the accuracy of the survey results and the consequent direct implications for the question of utility.

This paper will describe the design and underlying principles of the Census Bureau's evaluation program for the SIE. Because the report on the analysis is currently under review and revision and has not yet been submitted to the Congress, it is inappropriate to discuss publicly the estimates or conclusions of the evaluation program out of respect for the Congress. This paper will, however, present a statistical model that forms a component of the analysis, since this model is based entirely upon published data.

Considerations in the Design

The current government definition of poverty for statistical purposes is based principally upon money income and number of persons in the family, although the age and sex of the head, and the farm/non-farm status of the household are also included in the determination. Previous experience, particularly from the comparison of the Census with the Current Population Survey in 1970, has shown that the statistical measurement of poverty at the national level is sensitive to the choice of survey procedures. Furthermore, although differences in coverages and in definitions of household membership may contribute to differences between surveys, the available evidence pointed to problems in the collection of income data and in allocation for non-response as the primary driving force behind these differences. In general, obtaining accurate and complete income data from surveys and censuses has been problematic, but this difficulty has appeared most conspicuously in the poverty statistics.

These considerations suggested the particular formulation of the accuracy of the estimates in terms of their consistency among States. In other words, because the primary impetus for the survey was to obtain current estimates for use in an allocation formula, the survey results would be accurate for this purpose if they led to a correct allocation among States. As a first approximation, this would in turn be achieved by survey estimates that correctly represented each State's share of the national total of children aged 5-17 in poverty families, even if the national total was open to question.

Discussions between the Executive Branch and Congress led to the agreement for a specification of a coefficient of variation of 10 percent for each State's estimate of the number of children aged 5-17 in poverty families. Although the relation between this objective and the actual statistical reliability obtained by the SIE is an important question, the primary focus of the evaluation was to determine the possible effect of non-sampling errors in the State estimates.

Several previous evaluation programs to measure non-sampling error have been formulated in terms of the consistency of the respondents' answers over repetitions of the survey process or the uniformity of the interpretation and execution among interviewers. In planning the SIE evaluation, the analytic measures obtained from these other studies, "simple response variance" and "correlated" or "interviewer variance," were seen as at best tangentially related to the problem of non-sampling error in the SIE State estimates based upon the work of many interviewers and thousands of interviews. The perspective chosen instead was to determine directly the presence of systematic non-sampling errors affecting the SIE State estimates. This perspective led in turn to the decision to create an alternative survey process as a standard for comparison to the SIE. By conducting an alternative process of greater intensity than the SIE, the SIE survey estimates would be judged consistent within the limits of this standard if the more intense procedures would not change the allocation among States. Variance considerations forced this evaluation to be a reinterview of a subsample of the original sample, but conceptually the principles of analysis would have been similar if an entirely independent (but necessarily larger) evaluation survey had been conducted.

Because of an increasing legislative tendency to distribute public monies to subnational units according to need and to measure this need statistically, an increasing obligation has been placed upon the producers of these statistics to insure the consistency of the measurement process. The conceptual design for this evaluation may therefore serve as an example for future evaluations of this sort.

Design of the Reinterview

To create a standard for the evaluation of the SIE, two principles were followed: to obtain the critical information in the households selected for reinterview as independently as possible, and to increase the intensity of effort sufficiently to establish prime facie evidence that the

reinterview was indeed a valid standard for evaluation. As a consequence, the planning for the reinterview required an effort comparable to the planning for a new survey.

In the SIE and CPS, generally one person in a household served as the "household respondent" and provided all information, including on income, for all household members. A first specification in the reinterview design was to require self-response for all household members age 16 and over, even though call-backs were generally necessary to achieve this. Although hypothetical situations can be constructed where a self-respondent is less informed or cooperative than another household member, in general self-response was felt to be a better, although expensive, choice for the reinterview. (Some Census Bureau surveys, notably the National Crime Surveys, have required self-response when it has been judged that an increase in accuracy would justify a concommitant increase in cost.)

The second key feature of the design was the development of a new questionnaire. The new questionnaire incorporated a deliberate attempt to correct possible deficiencies of the SIE income section, which in turn had represented a minor modification of the corresponding CPS section. The CPS and SIE income sections record the data on recipiency and amounts in a FOSDIC-readable format. The questions on recipiency and amounts for each type of income follow each other in alternation. Although indicating all the necessary information to be obtained, this design provides neither the interviewer nor the respondent support in correctly determining amounts. It has also been suggested that the rapid succession of questions on recipiency leads respondents to say "no" more or less automatically, even when one of the types may actually have been received. Some CPS interviewers have also suggested asking all questions on recipiency first before any questions on amounts, since the latter are often the most sensitive issues. This would follow a general principle of questionnaire design, to precede the most sensitive questions with less sensitive ones.

The broad structure of the reinterview questionnaire is to establish in a "screening" section the recipiency by type of income in the context of a general review of possible income-related activities and situations during the year, and then to collect the amounts of the various types of income in "amounts" sections specifically designed for the particular types of income. For example, a respondent is asked in the screening section about all jobs held during the year. Later, in the amounts section for wages and salary, the respondent is questioned on each job separately. For each job, the respondent is first requested to consult a W-2 form for the information but, if unable or unwilling to do that, is allowed to provide an estimate if the respondent feels reasonably certain of the amount. If no figure can be obtained in this way, several alternative paths of questions assist the respondent in constructing an estimate based on an annual salary, an hourly or daily wage, or average amount paid in each paycheck. Consequently, the reinterviewer is directed through a series of questions that in general a resourceful interviewer might use with respondents requiring such help, but which is not provided by the CPS or SIE questionnaires.

A subsample of about 4.5 percent was selected for reinterview from the CPS and SIE samples. Approximately 2,000 and 6,000 reinterviews were obtained for the CPS and SIE, respectively. Stratification on the number of children aged 5-17 and the originally reported income was employed to reduce the sampling variance of the reinterview estimate of the number of children aged 5-17 in poverty families. In general, reinterviewers were provided only the information required to locate the original household, to insure the independence of the reinterview information. Subsequent edits in the field offices and later by computer identified a group of cases with significant discrepancies. This group was recontacted to assure the accuracy of the reinterview results. The preliminary analysis of these data has been completed. Because they form the basis for the evaluation to be reported to Congress, however, it is fitting to postpone the public discussion of the findings.

A Statistical Model for Children in Poverty

A standard statistical technique, linear regression, illustrates important aspects of the SIE estimates of children aged 5-17 in poverty families by State. Recent work in the application of this technique to survey estimates is due to Eugene Ericksen (1973, 1974). In general terms, sample estimates for the geographic units (counties, SMSA's or States) may be used as the dependent variable in a linear regression based upon symptomatic data gathered without sampling error for the same geographic units. The resulting predicted values are generally biased estimates of the population values for these geographic units, but in some applications they may possess considerably smaller average mean square errors than the sample estimates themselves. Furthermore, this technique allows the linear relationship between the symptomatic variables and the variable of interest to be determined directly from the current sample data, rather than from *a priori* reasoning or previous experience.

On the basis of this research, Census Bureau staff (Gordon Green and Robert Fay) studied the possible application of this technique to estimate the proportion of children aged 5-17 in poverty families for each State. The model was developed (in 1975) by attempting to fit the 1970 Census values for the percent of families in poverty by State on the basis of the corresponding 1960 Census results and other information. (Estimates of children aged 5-17 in poverty by State are not available from the 1960 Census.) These investigations favored a model based upon the census values and estimates of Per Capita Income (PCI) published in the Survey of Current Business by the Bureau of Economic Analysis. Sample estimates for the percent of children in poverty by State are fitted by a regression incorporating six independent variables: the constant term, the

census percent in poverty for the base year, and two variables derived from PCI figures for each of the base and current years. For each of two years of BEA data, the median, PCI_m, of the 51 State figures is determined and the variables

$$X_{j1} = \begin{cases} \ln (PCI_j / PCI_m) & \text{if } PCI_j > PCI_m \\ 0 & \text{otherwise} \end{cases}$$

$$X_{j2} = \begin{cases} 0 & \text{if } PCI_j > PCI_m \\ \ln (PCI_j / PCI_m) & \text{otherwise} \end{cases}$$

formed. The regression is weighted inversely proportional to the sampling variance of the sample estimates.

Ericksen's research included a possible approach to estimate the average mean squared error of the regression estimates. Basically, the sampling error of the sample estimates may be subtracted from the squared deviations between the sampled and fitted values to estimate the squared bias of the regression as the remainder. In this way, a current evaluation of the regression estimates may be obtained. The technique generally requires precise estimates of the sampling errors, however, and becomes ineffective in cases where the sampling errors completely dominate the biases of the regression.

Although the regression model has been fitted to the sample estimates of the percent of children 5-17 in poverty families by State from the CPS for all years subsequent to 1970, the sampling errors of the CPS estimates obviate any effective assessment of the fit. The SIE therefore affords the first such opportunity since the 1970 Census. Table 1 compares the 1970 Census estimates for 1969, the 1976 SIE estimates for 1975, and the model estimates based upon the SIE by State. The national poverty rates from the SIE and 1970 Census are virtually the same (14.5 percent vs. 14.8 percent), but there is a substantial redistribution of poverty among States. The changes estimated by the SIE since the 1970 Census correspond to an average of approximately 23 percent root mean square (r.m.s.) by State. Since the SIE estimates have an average c.v. of 10 percent, a real change of approximately 20 percent (r.m.s.) may be inferred ($23^2 \doteq 20^2 + 10^2$). On the other hand, the model estimates are within about 14 percent (r.m.s.) of the SIE values, leaving an unexplained bias of only about 10 percent r.m.s. ($14^2 \doteq 10^2 + 10^2$) between the model and SIE estimates. The model estimates therefore describe approximately 75 percent of the real change indicated by the SIE (10 percent r.m.s. vs. 20 percent r.m.s.).

The concurrence between the SIE and regression estimates has two important implications. The result reflects generally well upon the regression methodology: although not free from bias, the results closely resemble the actual survey outcome. Furthermore, if the regression estimates were to continue to explain 75 percent of the real change (a reasonable assumption according to the original research based on predicting the 1970 Census values from the 1960 Census), while the velocity of real change were also to continue at the rate for 1970-1976, it could be argued that the regression estimates based upon CPS data would be less biased estimates of the actual rates two or three years hence than the SIE rates for 1976.

The logic of the comparison may be reversed, however, and used to argue the face validity of the SIE survey estimates. Linear regression is a projection in the mathematical sense. In the application here, the model estimates are the projection of the 51 survey estimates onto a subspace of dimension 6. The residuals of the regression lie in a subspace of dimension 45. The residual subspace includes most of the sampling error in the SIE survey estimates, as well as the biases of the model estimates. If the SIE State estimates were subject to non-sampling errors, it might be assumed that the largest component of this error would also lie in the residual subspace. Therefore, the 14 percent r.m.s. difference between the model and survey estimates serves as an upper bound on the sum of the sampling error and this component of the non-sampling error. Even though 14 percent may be large, it still provides reassurance that the non-sampling errors of the survey State estimates are not extreme and arbitrary.

References

Ericksen, Eugene P. (1973), "A Method of Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas," Demography, 10, 137-60.

_____ (1974), "A Regression Method for Estimating Population Changes for Local Areas," Journal of the American Statistical Association, 69, 867-75.

Table 1. Percent of Children 5-17 Years Old in Poverty Families According to 1970 Census, SIE, and Regression Model

States by Division	1969 Estimate	1975 Estimates	
	Census	SIE	Regression Model
New England			
Maine	14.2	15.3	14.2
New Hampshire	7.7	10.3	10.5
Vermont	11.4	17.8	11.9
Massachusetts	8.4	9.3	10.6
Rhode Island	11.0	10.5	11.8
Connecticut	7.2	8.4	9.6
Middle Atlantic			
New York	12.2	13.1	13.8
New Jersey	8.7	11.6	10.2
Pennsylvania	10.6	12.6	10.9
East North Central			
Ohio	9.8	11.6	11.8
Indiana	9.0	9.6	10.8
Illinois	10.7	15.1	10.8
Michigan	9.1	11.3	11.2
Wisconsin	8.7	9.4	9.6
West North Central			
Minnesota	9.5	9.1	9.7
Iowa	9.8	7.9	8.2
Missouri	14.8	14.7	14.8
North Dakota	15.7	11.5	10.4
South Dakota	18.3	13.1	15.3
Nebraska	12.0	10.1	10.3
Kansas	11.5	8.6	10.2
South Atlantic			
Delaware	12.0	10.4	12.3
Maryland	11.5	10.7	11.2
District of Columbia	23.2	15.7	17.8
Virginia	18.2	13.7	15.0
West Virginia	24.3	18.9	18.2
North Carolina	24.0	17.8	20.2
South Carolina	29.1	23.9	23.4
Georgia	24.4	21.3	20.9
Florida	18.9	21.6	16.6
East South Central			
Kentucky	25.1	21.4	20.2
Tennessee	24.8	20.5	20.2
Alabama	29.5	15.9	23.1
Mississippi	41.5	32.6	32.2
West South Central			
Arkansas	31.6	21.4	23.8
Louisiana	30.1	22.9	23.8
Oklahoma	19.5	14.6	16.2
Texas	21.5	20.5	17.7
Mountain			
Montana	12.9	12.5	10.8
Idaho	12.0	11.0	10.5
Wyoming	11.2	8.6	8.2
Colorado	12.3	10.7	10.7
New Mexico	26.3	26.0	21.2
Arizona	17.5	16.8	16.1
Utah	10.0	8.0	9.4
Nevada	8.8	11.0	9.8
Pacific			
Washington	9.3	10.0	10.2
Oregon	10.3	8.4	10.2
California	12.1	13.8	12.5
Alaska	14.6	6.4	6.9
Hawaii	9.7	9.6	9.8

DISCUSSION

Sylvester J. Schieber, Social Security Administration

The Survey of Income and Education (SIE) described in this set of five papers may ultimately be judged a failure if the only judgmental criterion is the degree to which it fulfills its original goals. Ginsburg and Grob succinctly state the problems the Congress will have in determining the relative shares of Elementary and Secondary Education Act monies based on the estimates from the SIE of the number of children 5 to 17 years old living in poverty in each of the States. However, to judge the SIE on that basis alone would be an exercise in tunnel vision ignoring the SIE as a valuable resource to be exploited in the development of a wide range of research and policy evaluation. Already the SIE is being utilized extensively in the development of the current administration's welfare reform proposals now under consideration by the Congress. In fact the SIE is the most extensive body of data available for simulating the proposed welfare reforms and will certainly be the primary micro-data set used for that purpose during the evolution of the policy debate surrounding this particular proposal for the next couple of years. When one considers that the proposed welfare reforms may include a net increase in Federal expenditures of from five to ten billion dollars or more for 1980, then the expenditure of 14 million dollars to enlighten the policy debate seems more than worthwhile from almost any cost-benefit perspective. In addition to this use the SIE will provide information for analyzing various proposals and issues pertaining to tax structures, income transfer programs, related social programs, distributions of income and wealth and measures of economic and social well-being.

Despite my belief that current and potential benefits of this data base more than warrant the efforts and expense described by George Gray and Marvin Thompson in their paper I want to mention some reservations that I have with the Survey of Income and Education and its potential uses. Some of my reservations are peripheral to the SIE itself but pertinent to the more global process of data collection and analysis that have been addressed in this series of papers.

The first matter I wish to discuss is one of content and is raised because of current trends in analysis of micro-data sets of this sort. Specifically my reservations concern the attempt by the SIE to measure certain types of in-kind income. Ginsburg and Grob defined one of the specific areas of analysis for which these data were well suited as the measurement of the distribution of income and wealth. If in-kind benefits are to be included for low-income persons when deriving measures of relative shares of income then it is only fair to include them for higher income persons also. To do otherwise distorts the distributions being measured. Some analysts might contend that in-kind income accruing to the middle or upper income sectors of society is insignificant. I contend otherwise. Ask the man with a company car that can be used for personal purposes during non-business hours if it is of no value to him personally. Ask the corporate executive if his preferential stock options are worthless or middle level management personnel if their profit sharing and retirement packages are meaningless. Ask sales personnel if their prizes of vacations, cars, televisions, etc., won in sales competitions are of no value to them. Ask a very large portion of the working men and women in this society if their health benefits which are increasingly covering eye and dental care would be relinquished freely. What about life insurance, expense accounts, memberships in athletic or social clubs, clothing allowances, travel benefits, and educational benefits that commonly accrue as non-cash income to workers in our society? These kinds of income are of value, in many instances of significant value, and their receipt should at least be measured, even if their value cannot.

Also there is currently a popular trend to include in-kind income accruing to low-income persons in the calculations of the number of persons in poverty. I am not opposing the inclusion of in-kind benefits in income definitions, even though there are tremendous measurement problems. However, I am opposed to using the Orshansky poverty indices as currently defined as the relevant poverty thresholds if in-kind benefits are included. The Orshansky index brings together two separate food expenditure measures to define the poverty thresholds: (1) the cash expenditures needed to provide a family of given composition with a pre-defined level of nutrition; and (2) the ratio of total cash income to cash expenditures on food. That is, the poverty level income (PL) equals the product of cash food need (CFN) and cash income divided by food expenditures ($Income/Foodbill). Arithmetically that is:

$$PL = CFN \times (\$Income/Foodbill)$$

The cash food need component of this relationship is determined by measuring the costs of pre-defined bundles of food that meet certain nutritional requirements of families of given composition. The cash income/food expenditure component is an empirical measure derived originally for this purpose from the 1955 Food Consumption Survey. The important element to note in this relationship is that the income element used in defining poverty is cash income. If in-kind benefits are to be included in counting the poor then they should be included in the definition of poverty. If in-kind benefits are included in the income portion of the $Income/Foodbill ratio then the poverty thresholds for all classes of families would rise. Assuming the problems of measuring in-kind income are overcome and these benefits are included in both the definition and measurement of poverty, it is impossible, a priori, to estimate the net changes in the number of poor persons or in their characteristics from currently defined levels.

The second general area of concern regarding the SIE is the whole problem of error and how it potentially impacts on the ultimate analytical results which will be generated using the survey. The problems of error have been spelled out in at least four of the papers presented here. Ginsburg and Grob demonstrated the importance of error when they indicated that one standard error in the estimate of poor children in California could mean 10 million dollars in Elementary and Second-

TABLE 1. COMPARISON OF SIE AND MARCH 1976 CPS PERSON'S INCOME NONRESPONSE RATES

Type of Income	March 1976 CPS	SIE	CPS Rate / SIE Rate
Total	19.5	13.0	1.5
Wages or salary 1/	10.8	6.1	1.8
Nonfarm self-employment 1/	7.6	2.5	3.0
Farm self-employment 1/	7.2	2.1	3.4
Social Security or Railroad Retirement	11.2	2.6	4.3
Supplemental Security Income	10.1	1.5	6.7
Public Assistance or Welfare 2/	10.1	1.6	6.3
Interest from Savings Accounts	13.7	7.0	2.0
Dividends, rent, estates or trusts	11.7	3.7	3.2
Veterans' Payments, Unemployment Compensation, Workmen's Compensation	10.6	2.0	5.3
Private, Federal, Military, State and Local Pensions	10.5	1.9	5.5
Alimony and Child Support, Contributions from Persons not in the Household or any other Money Income	10.3	1.6	6.4

1/ Persons who did not work in 1975 who did not respond to the earnings questions were not considered nonrespondents for these items.
2/ Public assistance and welfare consists mainly of Aid to Families with Dependent Children and General Assistance.

ary Education monies for the State. Statistically, the extent of sampling error is relatively easy to identify and thus the implications of this sort of error can be measured. But sampling error is only one component of the total error included in any estimates from the SIE or similar data sets. Sampling error, in fact, may well comprise the smallest portion of total error in such estimates.

There are three papers in this set that deal specifically with various aspects of nonsampling error. Robert Fay describes a methodology for measuring nonsampling error but does not present findings, out of deference to the Congress, from actual tests of the model. He indicates that the perspective was to determine the presence of systematic nonsampling error. To do this a reinterview of "greater intensity" was conducted to serve as a benchmark against which SIE responses were judged. The stated goal of this process was to determine if Elementary and Secondary School Act monies would be allocated among the States in the same way using either the SIE or the reinterview as the basis for distribution. The basic assumption here, that the distribution of funds based on the SIE would be judged equitable if the more intense procedures would not change the allocation is not necessarily valid. A respondent or even a household comprising several respondents could have consciously provided corresponding misinformation on both the SIE and the more intensive reinterview. To the extent that nonsampling error was not random on the first interview it could potentially have been reinforced in the reinterview. Additionally, in the test described here, reliability of the survey can only be measured for those respondents providing complete information on both interview waves. This procedure itself may serve as a selection process for those respondents most willing and conscientious about providing correct information in the first place.

The Coder paper indicated considerable variance in nonresponse to the income items by State. For example, the income nonresponse rate in Connecticut (18.3 percent) was more than twice that in Arkansas (9.1 percent) or New Mexico (9.1 percent). Marks and Nisselson mention an upward bias in estimates of children in poverty families due to errors from "within household coverage." If this bias is accentuated by income nonresponse then the State variations in nonresponse rates could be quite important.

Marks and Nisselson are fairly specific in their discussion and estimation of noncoverage of households in the SIE. They estimate that between 6 and 11 percent of possible housing units were missed in the SIE sampling process depending on which reinterview subsample stratum of households was considered. When this noncoverage rate is combined with the income nonresponse rate of 13 percent discussed by Mr. Coder in his paper the nonsampling error problem is a matter for serious concern.

There is a corollary issue raised by Coder's paper regarding nonresponse to income questions on the March CPS. In his Table 1 he compares the SIE and March 1976 CPS person's income nonresponse rates. I have lifted the first two columns of that table and added a column indicating the magnitude of the differing nonresponse rates (i.e., Column 1 divided by Column 2) in Table 1.

The overall nonresponse to the whole set of income items was roughly one-third better on the SIE than the March CPS. Reducing overall income nonresponse by 33 percent is not insignificant. However, the improvement on an item-by-item basis varied considerably. For example the nonresponse rate for Social Security or Railroad Retirement on the March CPS was more than four times the rate on the SIE. For Supplemental Security Income the difference was nearly seven times, for Public

Assistance more than six times, etc.

The comparisons between the SIE and CPS clearly indicate that nonresponse to the income items on the March CPS could be reduced. It is widely held that more persistent pursuit of income information on the March CPS would actually result in lower response rates to labor force questions on subsequent waves of the CPS as Coder suggests. He also indicates, however, that there may be a clear advantage to collecting income information on the March CPS through personal versus telephone interviews. Thus the March CPS might benefit significantly from a more strenuous effort to reduce telephone interviewing. In addition, most of the income items from the March Supplement to the CPS are not directly related to the employment situation. In fact many of the recipients of welfare and pension income will have no attachment to the labor force during their tenure in a CPS rotation group. There is the possibility that more diligence in collecting non-wage income information might result in significantly improved data while having only a minimal impact on the gathering of subsequent labor force statistics. Thus it would seem there might be some reasonable trade-off between income data and slightly reduced labor force response rates. While the CPS was originally intended to gather labor force data, the March Supplement has become a major policy evaluation tool and thus the integrity of these data is of the utmost importance. Because of differences in income nonresponses 75 percent more income was allocated on the March 1976 CPS than on the SIE (i.e., 20 billion versus 12 billion dollars). Some effort should be made, at least on a limited basis, to improve the response rates to the income items of the March income supplement to the CPS. The SIE is proof that it can be done.

Another related issue, that does not follow directly from the SIE itself but is of critical importance in its utilization, is the problem of analytical error. This is an issue that has been widely ignored by the research community as well as the ultimate consumers of these data, the policy analysts and policy makers. Errors of this type arise because of mis-specifications of the issues being analyzed, because of the failure of the information available to fit the issues being tested and because of vagaries that exist in the computer software and simulation packages and procedures used to process the data. Hopefully, errors of mis-specification are caught by the professional community. Having data sets that are capable of fitting any analytical question is virtually impossible because the data sets usually precede the research problems. Potentially a very serious source of error in the analytical process, however, lies with the electronic data processing software. As the sophistication and complexity of the computer simulation and analytical software increases it is becoming more and more difficult for the analyst to be in control of the statistical and arithmetic operations actually performed. Increasingly the scenario is one of an analyst providing specifications for the task at hand, and the computer programmer converting those specifications into machine readable form. If there is any imprecise communication between the two it can result in error, potentially undetectable by either party. This potential for error is further compounded by the fact that in many instances there are large numbers of individuals who participate in this process in an evolutionary time frame. In the case of software performing standard statistical calculations, the results of newly created programs can be checked against previously existing ones. In the case of simulation systems of new or existing social programs this is not the case. To use these simulation programs is frequently quite simple. For example, assume we have an income maintenance simulation model: there is a requirement to specify the format of the input data elements (e.g., pertinent SIE data), a need to set certain exogenous parameters (e.g., tax rates, guarantee levels, unemployment rates, etc.) which are used in an iterative process of generation of a series of endogenous parameters (e.g., estimated asset levels, labor supply effects, etc.) that combine with all other information available to the system to generate caseload and cost estimates for a proposed income maintenance program. The estimated variables from each iteration of the model include stochastic error separate from the measurement error previously discussed. As the interactive process between analysts and the data processing machinery is simplified, the need for them to understand what specific calculations are actually performed in order to generate impressive and neatly formatted printed output is drastically reduced. The implications of the combined error factors are frequently overlooked.

The intention here is not to say that these simulations should not be performed. It is merely to point out that the problems of analytical error deserve our equal attention with those of sampling and other nonsampling error. We need to determine how these separate kinds of error combine and measure their implications on the estimates being generated.

The final issue addressed in this comment relates to the SIE as it fits into the time serial package of Census surveys. Grob and Ginsburg point out that there may not be comparability between poverty estimates generated from the SIE and the 1980 Census. If the SIE or similar surveys in the future are to bridge the gaps between the decennial Censuses, as Grob and Ginsburg suggest, then it would seem worthwhile to standardize key elements of the survey forms, data collecting procedures, etc., to guarantee that differences in measured phenomena are not the result of differences in measurement technique.

A PRECISE LIFE TABLE METHOD

John J. Hsieh, University of Toronto

1. INTRODUCTION

This paper provides a precise method of constructing abridged life tables. Such construction involves two problems: The main one is the estimation of the survival rate, ${}_nP_x = l(x+n)/l(x)$, from deaths registered during a given base period and populations enumerated or estimated at mid-years in each age interval; the secondary problem is the estimation of the stationary population ${}_nL_x$.

The estimation of the survival rate calls for the solution of certain equations which relate the observed age-specific death rate to the function underlying the age distribution in the stationary population on the one hand and the lifetime distribution in the stationary population on the other (Section 2). The solution of these equations could be regarded as an approximation of a dimensionless function by known dimensional functions. Keyfitz and Frauenthal (1975) solved such an equation and obtained an explicit functional relationship which approximated the survival rate in terms of age-specific death rates and mid-year populations, and which they showed are considerably more accurate than those derived by using the age distribution of the stationary population (Greville 1943). The main purpose of this paper is to provide a different set of explicit formulas (Section 3) which will be shown to be more accurate than Keyfitz-Frauenthal's and capable of removing the two defects inherent in their method (Section 4). A complete cubic spline obtained from consideration of the life-time distribution is used to compute the ${}_nL_x$ function and the result is shown to be more accurate than other existent methods (Section 5). Life tables so constructed are to be viewed as constructed at the midpoint of the base period. A detailed discussion of application of spline functions to life table construction, including the construction of complete life tables, is given in the more comprehensive paper, Hsieh (1977).

2. FUNDAMENTAL CONCEPTS AND EQUATIONS

We shall use a star superscript (*) to distinguish functions in the observed population from their corresponding functions in the stationary population. Let $l^*(x,t)$ be continuous having a continuous first order partial derivative with respect to age x, and represent the profile of the observed population pyramid at calendar time t (the unit of $l^*(x,t)$ is "persons per year") so that $l^*(x,t)dx$ is the number of individuals aged x to x+dx at time t and $l^*(x,t)$ dxdt is the number of individual-time units observed on the region dxdt. Let $\mu^*(x,t)$, which possesses similar regularity properties to $l^*(x,t)$ be the force of mortality at age x and calendar time t ($\mu^*(x,t)$ has unit "per year"). Then, the death rate ${}_{h}M^{t}_{n\,x}$ (whose unit is "per year") for the age interval [x,x+n) and the base period [t,t+h] (usually n=5 or 4 years, and h=3 or 1 year.) can be expressed as

$$ {}_{h}M^{t}_{n\,x} = \frac{\int_0^h \int_0^n l^*(x+v,t+u)\ \mu^*(x+v,t+u)\,dvdu}{\int_0^h \int_0^n l^*(x+v,t+u)\,dvdu} . \qquad (2.1) $$

The numerator in the above expression (whose unit is "persons") represents the number of individuals aged x to x+n who die during the base period [t,t+h] and is known from the death data. The denominator (whose unit is "person years") represents the individual-time units of exposure to the risk of death in the same age interval and base period and is unknown because the inner integral, which is the population between the ages x and x+n at any time point in the base period, is unknown except at midyears.

By definition, to construct a life table at the mid-period is to take the hazard function $\mu^*(x,t+h/2)$ of the observed population at this time point to be the hazard function $\mu(x)$ for the stationary population of the life table. Consider the time variable to be fixed at the midpoint t+h/2 of the base period and write

$$ {}_nP_x \equiv \int_x^{x+n} l^*(v,t+h/2)\,dv. $$

Then, (2.1), with the mid-period time point t+h/2 understood, can be written as

$$ \int_x^{x+n} l^*(v)\mu(v)\,dv = {}_nM_x\ {}_nP_x . \qquad (2.2) $$

The person-year integral in the denominator of (2.1) can be numerically integrated and expressed in terms of populations at mid-years (Hsieh, 1976). ${}_{h}M^{t}_{n\,x}$ can then be calculated from the observed data.

From the lifetime distribution theory or the pure death process we have for the lifelength X at midperiod,

$$ {}_nP_x \equiv \Pr\{X>x+n \mid X\geq x\} = \exp\{-\int_x^{x+n} \mu(v)\,dv\}. \qquad (2.3) $$

(2.3) expresses the survival rate ${}_nP_x$ in terms of the force of mortality $\mu(x)$ over the corresponding age interval in the stationary population.

3. CALCULATION OF THE SURVIVAL RATE

In (2.2) both quantities on the right hand side are known whereas both $l^*(v)$ and $\mu(v)$ on the left hand side are unknown functions. This equation can be regarded as an integral equation with $\mu(v)$ as the unknown function. Once $\mu(v)$ is solved for, ${}_nP_x$ is obtained from (2.3). Equation

(2.2) can be shown to be indeterminate (Hsieh, 1977). Thus, in order to uniquely determine $\mu(x)$ or the integral

$$\int_x^{x+n} \mu(v)\,dv$$

from (2.2), it is necessary to impose constraints either on $l^*(x)$ or on $\mu(x)$ alone or on both $l^*(x)$ and $\mu(x)$.

Aside from the standard regularity properties imposed on $l^*(x)$ in Section 2 for mathematical convenience, $l^*(x)$ also has its natural demographic properties: $l^*(x) \geq 0$ for $0 \leq x < \omega$, and $l^*(\omega)=0$, where ω is the maximum age. Now, letting

$$H(x) \equiv \int_x^{\omega} l^*(v)\,dv$$

and integrating by parts on the left hand side of (2.2) yields

$$\int_x^{x+n} l^*(v)\mu(v)\,dv = \mu(x)H(x) - \mu(x+n)H(x+n) + \int_x^{x+n} \mu'(v)H(v)\,dv, \qquad (3.1)$$

where the prime signifies derivative. A Taylor expansion on $H(v)$ about $v = x+n/2$ in linear terms gives

$$H(v) = H(x+n/2) - l^*(x+n/2)(v-x-n/2) + E(v), \qquad (3.2)$$

with error term

$$E(v) = -\int_{x+n/2}^{v} (v-y)\,l^{*\prime}(y)\,dy.$$

Entering (3.2) into the last term on the right hand side of (3.1) and carrying out the integration, we have

$$\int_x^{x+n} \mu'(v)H(v)\,dv = H(x+n/2)[\mu(x+n)-\mu(x)] - (n/2)\,l^*(x+n/2)[\mu(x+n)-\mu(x)] + l^*(x+n/2)\int_x^{x+n} \mu(v)\,dv + \int_x^{x+n} \mu'(v)E(v)\,dv. \qquad (3.3)$$

We shall now approximate the integral involving the error term $E(v)$. Using the integral expression for the error term in (3.2) in the last integral of (3.3), replacing one of the two functions which form the product in the integrands by its average value, and reversing the order of integration in the iterated integral gives:

$$\int_x^{x+n} E(v)\mu'(v)\,dv = -\int_x^{x+n} \left\{\int_{x+n/2}^{v} (v-y)\,l^{*\prime}(y)\,dy\right\}\mu'(v)\,dv$$

$$\simeq -(1/n)\left\{\int_x^{x+n}\int_{x+n/2}^{v} (v-y)\,l^{*\prime}(y)\,dy\,dv\right\} \times \left\{\int_x^{x+n} \mu'(v)\,dv\right\}$$

$$= -(1/n)\left[\int_{x+n/2}^{x+n}\left\{\int_y^{x+n} (v-y)\,dv\right\}l^{*\prime}(y)\,dy + \int_{x+n/2}^{x}\left\{\int_x^{y} (v-y)\,dv\right\}l^{*\prime}(y)\,dy\right][\mu(x+n)-\mu(x)]$$

$$\simeq (n/24)[l^*(x)-l^*(x+n)][\mu(x+n)-\mu(x)]. \qquad (3.4)$$

Combining (2.2), (2.3), (3.1), (3.3) and (3.4), and using $H(x)-H(x+n) \equiv {}_nP_x$, yields

$$ln\,{}_nP_x = -[1/l^*(x+n/2)]\Big[{}_nM_x\,{}_nP_x - \mu(x)\,{}_nP_x + (n/2)\,l^*(x+n/2)\{\mu(x+n)+\mu(x)\} - [H(x+n/2)-H(x+n) + (n/24)\{l^*(x)-l^*(x+n)\}] \times \{\mu(x+n)-\mu(x)\}\Big]. \qquad (3.5)$$

The approximations in (3.4) made use of the mathematical fact that the two functions

$$G_1(y)=\int_y^{x+n} (v-y)\,dv \quad \text{and} \quad G_2(y)=\int_x^{y} (v-y)\,dv$$

do not change sign in any age interval $[x,x+n)$, and the demographic fact that $\mu'(v)$ does not change sign except for one or three intervals where the relative minima or maximum of the $\mu(v)$ curve occur. In these intervals, however, the value of

$$\int_x^{x+n} \mu'(v)\,dv = \mu(x+n)-\mu(x)$$

is near zero and therefore the approximation has little effect on the result. At worst, the approximation may be regarded as taking the error $E(v)$ to be constant within each of these transition intervals.

Our next task is to approximate by numerical methods the unknown quantities that appear in (3.5) in terms of mid-period populations and death rates. We adopt the conventional division of the whole agespan for the abridged life table into 0, 1, 5, 10, ..., 85, 90, ω years, where ω is the maximum age to which any individual can live. The prsent proposed method may be used to advantage for wider age groups. However, data for single-year age intervals, even when available, are not reliable; if they were, many simple methods would produce life table functions about as accurate as those produced by sophisticated methods such as the present one.

Our life table method begins with age one and ends at the exact age marking the start of the terminal age interval (90 in this case). The precise method for the first year of life, because of gross underenumeration (and estimation) of infants, requires birth data and is therefore different from the method for ages beyond one (see Greville 1947). Life table functions for the terminal age interval, because of the unknown ω, are conventionally computed using the fact that $l(\omega)=0$ and the assumption that the age distribution of the observed population is identical with that of the stationary population.

The formula for computing n-year survival rate is as follows:

$$ln\,{}_np_x = -n\,{}_nM_x - n\,A_xB_x \,/\, {}_nP_x, \qquad (3.6)$$

where (i) for x = 1, n is 4 and

$$A_1 = (725\,{}_4P_1 - 418\,{}_5P_5 - 162\,{}_5P_{10})/12825,$$

$$B_1 = (475\,{}_4M_1 + 722\,{}_5M_5 - 114\,{}_5M_{10})/1083 - (365/31)D_m/(B-D_1+D_m), \text{ or}$$

$$B_1 = (-1120\,{}_4M_1 + 1444\,{}_5M_5 - 324\,{}_5M_{10})/855;$$

(ii) for x = 5, 10, ..., 75, n is 5 and

$$A_x = (9\,{}_5P_{x-5} - 3\,{}_5P_x - 5\,{}_5P_{x+5} - {}_5P_{x+10})/192,$$

$$B_x = (-3\,{}_5M_{x-5} - 3\,{}_5M_x + 7\,{}_5M_{x+5} - {}_5M_{x+10})/8;$$

and (iii) for x = 80, 85, n is 5 and

$$A_x = ({}_5P_{x-10} + 2\,{}_5P_{x-5} - 3\,{}_5P_x)/48,$$

$$B_x = ({}_5M_{x-10} - 4\,{}_5M_{x-5} + 3\,{}_5M_x)/2.$$

Formula (3.6) is obtained by using collocation polynomials, ${}_nP_x \equiv H(x)-H(x+n)$ and the approximations ${}_nP_x = n\ell^*(x+n/2)$ and ${}_nM_x = \mu(x+n/2)$ in equation (3.5). For age intervals other than the first, the following general form of Newton's formulas with various chosen values of j, r and s,

$$f_{x+(j+r)n} = \sum_{i=0}^{s} \frac{r(r-1)\ldots(r-i+1)}{i!}\,\Delta^i f_{x+jn} + E_t,$$

where $\Delta^i f_x$ designates the ith forward difference of f_x and E_t denotes the truncation error, were used to express the unknown functions in (3.5) as linear combinations of mid-period populations and death rates. Because unequal age intervals were involved, Lagrange's formulas for collocation polynomials were employed for the first age interval (x=1, n=4). Also, the abrupt bend of the $\mu(x)$ curve around age one renders it inappropiate, except for countries with very low infant mortality, to extrapolate $\mu(1)$ in terms of death rates in succeeding age intervals. Since $\ell(x)$ is convex at age one, $\mu(1)$ is closely estimated by the ratio of the conditional probability of dying in the 12th month of life to the length of the month:

$$\mu(1) = (365/31)D_m/[B - D_1 + D_m], \quad (3.7)$$

where D_m, D_1 and B, respectively, denote the number of deaths in the 12th month of life, the number of deaths under one year, and the number of births, all during the base period. The data for D_m and D_1 are given for various countries in the 1974 U.N. Demographic Yearbook. Utilization of (3.7) leads to the alternative expression for B_1 given in (3.6).

4. COMPARISON OF ACCURACY

Keyfitz and Frauenthal (1975) showed that their life table method is more accurate than other ones. In this section we emphasize comparisons between the new method and the Keyfitz-Frauenthal (denoted henceforth as "K-F") method.

To effect a precise comparison of accuracy, we use the test proposed by Keyfitz and Frauenthal (1975) which assumes both functions $\ell^*(x)$ and $\ell(x)$ to be known, where $\ell(x)$ is the number of survivors to age x out of $\ell(0)$ births in the life table so that $\ell(x+n)=\ell(x)\,{}_nP_x$. Adopting K-F stable population profile $\ell^*(x)$ and Makeham's graduation formula for $\ell(x)$,

$$\ell^*(x) = 10^6[1-\exp(x/100-1)] \quad (4.1)$$

and $$ln\,\ell(x) = ln\,\ell(0) + x\,ln\,s + (c^x-1)\,ln\,g\,, \quad (4.2)$$

where s=.999859, g=.999743 and c=1.109887; $\mu(x)$, $\mu'(x)$, $\ell^{*\prime}(x)$, ${}_nM_x$ and ${}_nP_x$ are computed using (2.3) and (2.2).

Next, the new formula (3.6) and the formulas for the following abridged life table methods:

Greville:
(1943)

$$ln\,{}_nP_x = -n\,{}_nM_x - n^2\,{}_nM_x({}_nM_{x+n} - {}_nM_{x-n})/24, \quad (4.4)$$

Reed and Merrell:
(1939)

$$ln\,{}_nP_x = -n\,{}_nM_x - .008n^3\,{}_nM_x^2, \quad (4.5)$$

Keyfitz and Frauenthal:
(1975)

$$ln\,{}_nP_x = -n\,{}_nM_x + n({}_nP_{x+n} - {}_nP_{x-n})({}_nM_{x+n} - {}_nM_{x-n})/(48\,{}_nP_x) \quad (4.6)$$

are applied to the synthetic ${}_nM_x$ and ${}_nP_x$ to reproduce the life table $\ell(x)$. Since K-F formula (4.6) cannot be used for computing ${}_nP_x$ for the initial interval [0,5), the simple formula $ln\,{}_nP_x=-n\,{}_nM_x$ obtained from (2.2) and (2.3) by assuming constant force of mortality within this age interval, is used to compute $\ell(5)$ for all life table methods. The results are shown in Table 1. The cumulative absolute errors are found to be 4.55 for the new formula (3.6), 41.71 for the K-F formula (4.6), 825.66 for the Reed and Merrell formula (4.5) and 996.18 for the Greville formula (4.4).

The principal advantage of the new method over the K-F method is that the latter requires estimation of $\ell^{*\prime}(x)$ and $\mu'(x)$ while the former requires estimation of $\ell^*(x)$, $\mu(x)$ and

$$\int_{x+n/2}^{x+n} \ell^*(v)\,dv.$$

The well known fact that approximate derivatives

Table 1. Comparison of Exact Makeham $l(x)$ with Results of Four Approximate Life Table Methods

Age x	exact	Hsieh (3.6)	Keyfitz & Frauenthal (4.6)	Reed & Merrell (4.5)	Greville (4.4)
0	100000	100000	100000	100000	100000
5	99912	99912	99912	99912	99912
10	99812	99812	99812	99812	99812
15	99692	99692	99692	99692	99692
20	99538	99538	99538	99538	99538
25	99327	99327	99327	99328	99328
30	99021	99021	99021	99022	99022
35	98555	98555	98555	98556	98556
40	97822	97822	97821	97825	97825
45	96646	96646	96646	96652	96652
50	94744	94744	94743	94753	94753
55	91668	91668	91667	91684	91683
60	86754	86754	86752	86778	86776
65	79104	79104	79101	79134	79129
70	67747	67747	67741	67767	67754
75	52207	52208	52200	52176	52148
80	33679	33681	33670	33531	33481
85	16105	16107	16096	15828	15762
90	4651	4651	4647	4394	4346

Table 2. Comparison of Exact ${}_nL_x$ Computed from Makeham $l(x)$ with Results of Four Approximate Integration Methods

Age x	exact	(5.1)	Keyfitz & Frauenthal (8.2)	Polynomial (8.3)	Simple ratio (8.4)
1	399792	399790	399790	399790	399860
5	499316	499316	499316	499316	499520
10	498770	498770	498771	498771	499079
15	498092	498092	498093	498093	498532
20	497193	497193	497194	497194	497776
25	495920	495921	495923	495923	496648
30	494023	494024	494027	494027	494882
35	491082	491082	491089	491088	492049
40	486402	486403	486414	486413	487441
45	478855	478856	478873	478870	479909
50	466638	466640	466667	466656	467615
55	446992	446994	447038	447008	447748
60	415995	415996	416064	415988	416316
65	368844	368839	368942	368771	368461
70	301562	301546	301685	301379	300247
75	215361	215324	215488	215141	213169
80	122917	122916	123014	122987	120469
85	48574	48619	48613	49077	46870
Cumulative absolute error		114	677	1134	16347

Table 3. Abridged Life Table for Male Population: Canada, 1970-72

Age Group x- (1)	$_nP_x$ (2)	$_nD_x$ (3)	$_nM_x$ (4)	$_nq_x$ (5)	$\ell(x)$ (6)	$_nd_x$ (7)	$_nL_x$ (8)	T(x) (9)	e(x) (10)
Under 1	182195	11173	0.020441	0.020022	100000	2002	98226	6933697	69.337
1-4	747410	2119	0.000945	0.003800	97998	372	391106	6835470	69.751
5-9	1152430	1913	0.000553	0.002843	97625	278	487398	6444365	66.011
10-14	1181450	1837	0.000518	0.002595	97348	253	486205	5956967	61.193
15-19	1074430	4697	0.001457	0.007292	97095	708	483891	5470762	56.344
20-24	941775	5266	0.001864	0.009267	96387	893	479666	4986871	51.738
25-29	800710	3556	0.001480	0.007369	95494	704	475669	4507205	47.199
30-34	660875	3287	0.001658	0.008271	94790	784	472058	4031536	42.531
35-39	645045	4243	0.002193	0.010911	94006	1026	467645	3559478	37.864
40-44	640765	6886	0.003582	0.017771	92981	1652	461080	3091833	33.252
45-49	613415	10406	0.005655	0.027980	91328	2555	450757	2630753	28.805
50-54	518895	14562	0.009354	0.045945	88773	4079	434378	2179996	24.557
55-59	472415	20730	0.014627	0.070894	84694	6004	409427	1745617	20.611
60-64	381690	26571	0.023205	0.110425	78690	8689	372915	1336191	16.980
65-69	296050	31482	0.035447	0.163899	70001	11473	322435	963276	13.761
70-74	205575	32751	0.053105	0.235759	58528	13798	258880	640840	10.949
75-79	139995	33145	0.078919	0.330026	44729	14762	186786	381961	8.539
80-84	85680	30650	0.119242	0.456339	29967	13675	114579	195175	6.513
85-89	40625	21181	0.173793	0.592992	16292	9661	55166	80595	4.947
90+	13940	10905	0.260760	1.000000	6631	6631	25430	25430	3.835

obtained from collocation polynomials conglomerate much larger errors than do approximations of functions and their integrals is reflected in the ample difference (41.71 versus 4.55) in the cumulative absolute error between the two life table methods based on the results of Table 1. With the transition from synthetic to real data, the K-F method would suffer still greater loss in accuracy than the new method. This is because the analytic curve used in the test may be close to the true curve and yet the two curves still may have very different slopes. Thus, for age distributions with dents and bulges such as those resulting from the two World Wars, the estimated values of $l^{*'}(x)$ in age intervals adjacent to where the dents and bulges occur could differ vastly from the true values of $l^{*'}(x)$.

Another advantage of the present method over the K-F method regards coverage of the agespan. Since the K-F method requires three consecutive age intervals of equal length to calculate the survival rate for the central age interval, the K-F formula (4.6) cannot be used to compute survival rates for both the first age interval, either [0,5) or [1,5), and the last age interval [85,90). On the other hand, with no problem of estimation of slopes, the new formula (3.6) covers these two intervals just as well as other age intervals.

5. COMPUTATION OF STATIONARY POPULATION

With values of the survivorship function $l(x)$ available at the age points x=1, 5, 10, ..., 90, we now turn to the problem of computing stationary populations

$$ {}_nL_x = \int_0^n l(x+v)\,dv. $$

Various methods of approximating this integral exist in the literature with varying degrees of accuracy (see Table 2). We use the method of splines to approximate the integral

$$ L_i \equiv {}_{n_i}L_{x_i},\ i=0,1,\ldots,k, $$

by the formula:

$$ L_i = n_{i+1}(l_i+l_{i+1})/2 + n_{i+1}^2(s_i-s_{i+1})/12, \quad (5.1) $$

where $l_i \equiv l(x_i)$ and the slopes $\{s_i\}$ are to be determined by solving the following system of k-1 equations for cubic splines:

$$ n_{i+1}s_{i-1} + 2(n_{i+1}+n_i)s_i + n_i s_{i+1} $$
$$ = 3\left[\frac{n_i}{n_{i+1}}(l_{i+1} - l_i) + \frac{n_{i+1}}{n_i}(l_i-l_{i-1})\right], \quad (5.2) $$

$$ \text{for } i=1,2,\ldots,k-1, $$

with the two boundary conditions:

(1) the first endslope

$$ s_0 = l'(1) = -(365/31)l(1)D_m/[B-D_1+D_m], \quad (5.3a) $$

(2) the last endslope

$$ s_k = l'(90) = -l(90)({}_5M_{85})^{3/2}({}_5M_{80})^{-1/2}. \quad (5.3b) $$

The above boundary conditions define a complete cubic spline and are obtained on the basis of properties of life table functions. For further details see Hsieh (1977).

In Table 2 we compare four methods of computing approximate values of L_i, using the same set of data $\{x_i, l_i\}$ taken from Makeham curve (4.2). The exact values of L_i are obtained by integrating $l(x)$ in (4.2) from x_i to x_{i+1}.

The other three methods are:

Keyfitz and Frauenthal:

$$ L_i = \frac{n(l_i-l_{i+1})}{\ln l_i - \ln l_{i+1}}\left[1+n(M_{i+1}-M_{i-1})/24\right] \quad (5.4) $$

Polynomial (cubic):

$$ L_i = (13/24)n_i(l_{i+1}+l_i) - n_i(l_{i+2}+l_{i-1})/24 \quad (5.5) $$

Simple ratio:

$$ L_i = (l_i-l_{i+1})/M_i. \quad (5.6) $$

The age specific death rates

$$ M_i \equiv {}_{n_i}M_{x_i} $$

in (5.4) and (5.6) are computed from (2.2) using (4.1) and (4.2).

The results obtained from (5.1) generate a cumulative absolute error of 114, as compared with 677, 1134 and 16347 for formulas (5.4),(5.5) and (5.6) respectively.

To illustrate the present proposed method, formulas (3.6) and (5.1) are used to construct an abridged life table for the 1970-72 Canadian male population as shown in Table 3.

ACKNOWLEDGEMENT

Grateful acknowledgement is made to Douglas M. Okamoto for his assistance in performing the tests and in computing the three tables.

REFERENCES

1. Greville, T.N.E. (1943),"Short Methods of Constructing Abridged Life Tables." Rec. Am. Inst. of Actuaries 32: 29-42.
2. Greville, T.N.E. (1947), United States Life Tables and Actuarial Tables, 1939-1941. Washington, D.C.: National Office of Vital Statistics 117-122.
3. Hsieh, J.J. (1976), "Methods of Computing Person Years." Presented at Annual Joint Statistical Meetings in Boston.
4. Hsieh, J.J. (1977). "A Method of Life Table Construction and Spline Interpolation." Submitted to the Journal of the American Statistical Association.
5. Keyfitz, N. and Frauenthal, J. (1975), "An Improved Life Table Method." Biometrics 31: 889-8
6. Reed, L.J. and Merrell, M. (1939), "A Short Method for Constructing An Abridged Life Table." Am. J. Hygiene 30: 33-62.

ANALYSIS OF FERTILITY BY INCREMENT-DECREMENT LIFE TABLES

C.M. Suchindran, N.K. Namboodiri, and K.K. West, University of North Carolina

INTRODUCTION

What should be the explanandum in fertility research is a question that has attracted some attention in the literature recently. It cannot be said, however, that the question has been answered to the satisfaction of all. Many studies based on cross-sectional data continue to use as explananda summary measures that are proxies to a complete reproductive history. Examples are such measures as completed family size, and expected, desired, or ideal family size. Several writers have expressed the view that it is more logical to regard the reproduction process as a contingent sequence of events and that it is advisable to treat as explananda the probability and timing of each event in the sequence (see e.g., Mishler and Westoff, 1955; Namboodiri, 1972, 1974; and Ryder, 1975). In this view the occurrence of each event in the sequence is considered necessary but not sufficient for the occurrence of subsequent events. The arrival of the first baby, for example, is a prerequisite but not a guarantee for the conception of a second child. Once we recognize that it is fruitful to think of the reproductive process as a contingent sequence, it becomes interesting to ask: How does one describe the process in terms of meaningful fertility measures? In this paper we shall show that the reproductive process conceived as a contingent sequence can be conveniently described by means of an increment-decrement table. In the immediately following section we describe this procedure, using for illustration data from the 1965 U.S. National Fertility Study.

For technical expositions of the increment-decrement tables, reference may be made to Jordan (1967) and Schoen (1975), and for an application of the technique in the analysis of marriage history, see Schoen and Nelson (1974).

AN ILLUSTRATION

The data used in this section are, as stated already, from the 1965 U.S. National Fertility Study. Reference may be made to Ryder and Westoff (1971), for a detailed description of the sample design used in that study. In brief, the universe represented by the sample consisted of currently married women born since July 1, 1910, living, with their husbands, within coterminous United States, and able to participate in an English language interview.

For the present purpose, we shall use only a part of this sample. We shall confine attention to currently married women, with no history of marital dissolution and no premarital or multiple births. Our first analysis will be confined to women married at least 9 years.

The reproductive history of women in the subsample up to the fourth birth is summarized in Table 1. The tabulation was stopped with the fourth birth because the number of women with five or more births was too small to provide reliable information about the later phases of the reproductive process. (It would have been desirable to stratify these women by age at marriage or into birth cohorts and consider each stratum separately but the smallness of the sample size prevented us from doing this.)

To facilitate a formal description of the relationships between the figures in Table 1, let us introduce the following notation: Let

N_x^i = number of women at parity i at the completion of x years after marriage (e.g., N_1^0 = 1,807, in Table 1);

D_x^i = number of women who move from parity i to parity i+1 during the xth year after marriage (e.g., D_2^0 = 1,807 - 1,033 = 774, in Table 1);

and W_x^i = number of women who are reported to be at parity i and have been married for only x years as of the survey date (e.g., W_9^0 = 6 in Table 1).

It can be seen that the following relationship prevails between the figures in columns 2, 3 and 4 of Table 1:

$$N_x^0 = N_{x-1}^0 - D_{x-1}^0 - W_{x-1}^0$$

Thus, for x = 10, 199 = 213 - 8 - 6. In column 2, we thus see only decrements and no increments. (Had we incorporated marital disruption into the picture, the situation would have been different.) When we move to columns 5, 8 or 11, we see both increments and decrements. In column 5, the successive numbers are interrelated in the following manner:

$$N_x^1 = N_{x-1}^1 - D_{x-1}^1 + D_{x-1}^0 - W_{x-1}^1$$

Thus, for x = 10,

$$336 = 400 - 59 + 8 - 13.$$

Similarly, in column 8, we have

$$N_x^2 = N_{x-1}^2 - D_{x-1}^2 + D_{x-1}^1 - W_{x-1}^2$$

and so on.

The probability of moving from parity 0 to parity 1 (i.e., of having the first birth) in the xth year after marriage can be approximately calculated as

$$q_x^0 = \frac{D_x^0}{N_x^0 - \frac{1}{2}W_x^0}$$

and similarly the probability of moving from parity 1 to parity 2 (i.e., of having the second birth) in the xth year after marriage can be approximately obtained as

$$q_x^1 = \frac{D_x^1}{N_x^1 + \frac{1}{2}D_x^0 - \frac{1}{2}W_x^1}$$

and, in general,

$$q_x^i = \frac{D_x^i}{N_x^i + \frac{1}{2}D_x^{i-1} - \frac{1}{2}W_x^i} \qquad i \geq 1. \qquad (1)$$

The structure of these formulae can be easily understood when it is realized that what we are calculating is the frequency of occurrence of an ith birth per person-year of exposure. We assume that each of those who move into parity i-1 during a given year is exposed one-half year, on average, to the risk of having an ith birth. Similarly, we assume that each of those reported to be at

parity i-1 at the date of the survey has been exposed one-half year, on average, to the risk of having an ith birth before the survey date.

The q_x^i values calculated using the formulae just described are shown in Table 2. On the basis of these figures the reproductive history of a hypothetical cohort of 100,000 women has been constructed. This history is also reported in Table 2. Note that

l_x^i denotes the number of women of the original cohort (of 100,000) who reach parity i at the completion of x years after marriage,

d_x^i denotes the number of women who move from parity i to parity i+1 during the xth year after marriage,

and q_x^i denotes the conditional probability of moving from parity i to parity i+1 during the xth year after marriage.

It is easily seen that

$$d_x^i = q_x^i (l_x^i + \tfrac{1}{2}d_x^{i-1}) \qquad (2)$$

$$l_{x+1}^0 = l_x^0 - d_x^0, \text{ and}$$

$$l_{x+1}^i = l_x^i - d_x^i + d_x^{i-1}, \quad i=1, 2, \ldots. \qquad (3)$$

From Table 2 we can calculate a number of summary measures indicating the nature of the sequential process that reproduction is. A few of these measures are described below.

1. Parity Progression Ratio

The sum of the d_x^0 column in Table 2 represents the number of women in the original cohort (of 100,000) who ever move to parity 1. Similarly, the sum of the d_x^1 column represents the number of women who ever move from parity 1 to parity 2, and so on. From these column totals, we can calculate a sequence of parity progression ratios. Thus, for the progression from parity 0 to parity 1, we have

$$PP_{0,1} = \frac{\Sigma d_x^0}{100{,}000}$$

and for the progression from parity i to parity i+1 we have

$$PP_{i,i+1} = \frac{\Sigma d_x^i}{\Sigma d_x^{i-1}}, \quad i=1, 2, \ldots. \qquad (4)$$

The figures calculated in this fashion from Table 2 are: $PP_{0,1}$ = 93,897/100,000 = .9390; $PP_{1,2}$ = 84,113/93,897 = .8958; $PP_{2,3}$ = 58,689/84,113 = .6977; $PP_{3,4}$ = 37,161/58,689 = .6332. These figures indicate that almost 94 percent of the original (hypothetical) cohort bear at least one child; that among those who bear at least one child, 90 percent bear at least two children; that among those who bear at least two children, 70 percent go on to have at least three children; and that among those who attain parity three, 63 percent move on to parity four.

2. Mean Interval between Marriage and Successive Births

As stated already, the d_x^i column in Table 2 gives the numbers of women who make the transition from parity i to parity i+1 during the xth year after marriage. Assuming that these movements from parity 0 to parity 1 are evenly distributed within each year after marriage, we can calculate the mean interval between marriage and successive births from column 1 and 3 of Table 2 using the formula

$$AI_{0,i} = \frac{\sum_x (x + \tfrac{1}{2}d_x^{i-1})}{\sum_x d_x^{i-1}}, \quad i = 1, 2, \ldots \qquad (5)$$

The figures thus calculated from Table 2 are shown below: $AI_{0,1}$ = 2.40 years; $AI_{0,2}$ = 5.28 years; $AI_{0,3}$ = 8.03 years; and $AI_{0,4}$ = 10.47 years. It should be noted that the mean interval $AI_{0,1}$ represents the experience of all those who make the transition from parity 0 to parity 1, irrespective of what happens to them beyond parity 1. Some of these women may or may not make the transition to higher parities. Similarly, the mean interval $AI_{0,2}$ represents the experience of all those and only those who move from parity 1 to parity 2. Because of these changes in the bases, it is not strictly valid to interpret the difference

$$AI_{0,i+1} - AI_{0,i}$$

as an inter-birth interval. One way to avoid this difficulty is to include in Table 2 only women who had at least, say, 4 births; then the bases of $AI_{0,i}$ will be the same for i = 1, 2, 3, and 4.

3. Average Parity Attained within a Given Interval after Marriage

From Table 2, it is possible to calculate the average number of births occurring to the hypothetical cohort of 100,000 during any specific interval after marriage. Suppose, for example, we want to calculate the average number of births occurring during the first three years after marriage. This can be obtained by adding the numbers in the d_x^i columns for all i and for x = 0, 1, and 2, and dividing the sum thus obtained by l_0 (i.e., 100,000). The figure thus calculated from Table 2 is {(26,032 + 31,682 + 12,485) + (3,807 + 15,431) + 696}/100,000 = 0.9013. Note that the sum (26,032 + 31,682 + 12,485) represents the number of first births during the first three years after marriage, the sum (3,807 + 15,431), the number of second births during the same period, and 696, the number of third births during the period. A general formula for the purpose is

$$AP_{0,j} = (1/l_0) \sum_{x=0}^{x=j-1} \sum_{i=0} d_x^i \qquad (6)$$

where $AP_{0,j}$ stands for the average parity attained during the first j years after marriage. One can similarly calculate the average number of births occurring in any specific interval after marriage, e.g., between the fifth and tenth year after marriage.

4. Conditional Probability of Transition to Higher Parities

From Table 2, one can calculate conditional probabilities of the following types:

A. Given that a woman is at parity 0 when she completes 5 years after marriage (i.e., when she just starts her sixth year after marriage), what is the probability that she will bear her first child within the year? From Table 2, the required

probability is easily seen to be 0.18063.

B. Given that a woman has just reached her sixth year after marriage and is still childless, what is the probability that she will bear her first child sometime during the next 5 years? From the 1^0_x column of Table 2, we notice that 15,639 women reached the sixth year after marriage and are still childless. After 5 more years their number decreases to 8,390. Hence, the required probability is simply obtained as

$$\frac{15,637 - 8,387}{15,637} = 0.4636$$

which means that just less than 50 percent of these 15,637 women are likely to bear at least one child within the next five years.

Recall that in preparing Tables 1 and 2, birth events were related to the x variable, duration of marriage. One can use instead of duration of marriage the wife's age. This is illustrated in Tables 3 and 4. Table 3 pertains to all white women in the 1965 NSF sample who had no history of marital dissolution and no experience of premarital or multiple births. Note particularly that unlike in Table 1, where only women who had been married 9 or more years by the survey date were included, no parallel restriction with respect to age was imposed in the construction of Tables 3 and 4.

From Table 4 one can calculate a number of summary measures of the kind mentioned earlier. But summary measures taken singly or in combination do not portray the details of the information contained in the q^i_x sequences. So it is natural to ask: If summary measures tend to sacrifice information, why not work with the q^i_x sequences themselves displayed in tabular form? There are two major difficulties in doing this. First, the q^i_x sequences contain too many numbers to digest, and this is the reason, in the first place, why one tries to get summary measures. Second, and more important, the q^i_x sequences show a good deal of irregularities (see Figures 1 to 4). This problem becomes more serious when interest centers in comparative analysis of q^i_x values for population subgroups (e.g., religious and socio-economic classes), for in such situations, due to small numbers, sampling errors associated with the observed q^i_x values will be large. (Another reason for irregularities in the q^i_x sequences may be measurement errors.)

Demographers are familiar with several procedures for removing irregularities in observed rates and estimated risks. Among these are (1) curve fitting, and (2) grouping. Application of these two techniques in the present case are briefly discussed below.

Curve fitting: The Hadwiger function (see below) was found to give better fit to the q^i_x sequences (in Table 4) than some of the other (e.g., beta and gamma) functions often used in this type of exercises. In the present case the Hadwiger function has the form

$$q^i_x = \frac{RH}{2T\sqrt{\pi}} \left(\frac{T}{x-i}\right)^{3/2} \exp\left[-H^2\left(\frac{T}{x-i} + \frac{x-i}{T} - 2\right)\right]$$

where i stands for parity, x for age, and R, H, and T are parameters to be estimated (e.g., using methods for fitting nonlinear regression). The estimated values of R, H, and T for the data in Table 4 are shown below. It would have been nice if we could give meaningful physical interpretations to these parameter estimates. Unfortunately we have not been able to do this.

Grouping: This technique involves aggregating persons (women in the present case) and events (births or withdrawals) on the x variable (e.g., into age groups x to x + n) and then recovering from the information available for the aggregated data estimates of q^i_x values for single years. The data presented in Table 3 are reproduced in the aggregated form in Table 5.

Karup-King multipliers were applied to the numbers in columns of Table 5 to obtain the numbers of persons at pivotal ages 15, 20, 25, . . ., as well as withdrawals and births at these ages. From these pivotal numbers, q^i_{15}, q^i_{20}, . . . were calculated using formula (1). Karup-King multipliers were then applied to these pivotal q values to obtain q^i_x for all x. This procedure is now being examined for its robustness as different criteria for aggregation and different pivotal ages are used.

So far our attention in this paper has been devoted to constructing complete increment-decrement life tables. For many purposes, however, abridged life tables would be sufficient. To construct abridged life tables we proceed like this. From aggregated data shown in Table 5, we calculate ${}_5q^i_x$ values according to the following formula:

$${}_5q^i_x = \frac{{}_5D^i_x}{{}_5N^i_x + a^{i-1}_x \, {}_5D^{i-1}_x - \frac{1}{2}{}_5W^i_x} \qquad (8)$$

where ${}_5D^i_x$ = number of ith parity births in the age groups (x, x+5), ${}_5N^i_x$ = number of women remaining at the ith parity at the beginning of the 5-year interval (x, x+5), ${}_5W^i_x$ = number of women in the age group (x, x+5) who are withdrawn (from observation) when they are at parity i, and a = average fraction of the interval (x, x+5) spent at ith parity by women before moving to parity i+1. How a varies by age and parity remains to be investigated. One set of estimates of a obtained from the 1965 NSF are reported in Table 6. Table 6 contains estimated ${}_5q^i_x$ values obtained using these a's. In constructing Table 6 we found it more appropriate to use instead of (8) a modified version of Greville's formula (see Shryock and Siegel, 1972, pp. 444) for age group 15-19.

References

Jordan, C.W. (1967). Life Contingencies. The Society of Actuaries. Chicago, Illinois.

Mishler, G. and Westoff, C.F. (1955). A Proposal for Research on Social Psychological Factors Affecting Fertility: Concepts and Hypotheses. Pp. 121-150 in Milbank Memorial Fund (ed.), Current Research in Human Fertility. New York: Milbank Memorial Fund .

Namboodiri, N.K. (1972). Some Observations on the Economic Framework for Fertility Analysis. Population Studies 26: 185-206.

Namboodiri, N.K. (1974). Which Couples at Given Parities Expect to Have Additional Births? Demography 11: 45-56.

Ryder, N.B. (1975). Fertility Measurement Through Cross-Sectional Surveys. Social Forces 54: 7-35.

Ryder, N.B. and Westoff, C.F. (1971). Reproduction in the United States, 1965. Princeton:

Princeton University Press.
Schoen, R. (1975). Constructing Increment-Decrement Life Tables. Demography 12: 313-324.
Schoen, R. and Nelson, V.E. (1974). Marriage, Divorce, and Mortality: A Life Table Analysis. Demography 11: 267-290.
Shryock, H.S. and Seigel, J.S. (1973). The Methods and Materials of Demography. U.S. Department of Commerce, Washington, D.C.

TABLE 1

Observed Timing of Transition from One Parity to the Next: 2,443 Selected White Women (Married for 9 Years or More with No History of Marital Dissolution or Premarital Births): 1965 U.S. National Fertility Study

Marital Duration (in completed years)	Parity 0			Parity 1			Parity 2			Parity 3		
	Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date	Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date	Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date	Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date
1	2	3	4	5	6	7	8	9	10	11	12	13
0	2,443	636										
1	1,807	774		636	93							
2	1,033	305		1,317	377		93	17				
2	728	210		1,245	406		453	101		17	2	
4	518	136		1,049	319		758	194		116	35	
5	382	69		866	243		883	187		275	61	
6	313	50		692	189		939	177		401	93	
7	263	31		553	107		951	163		485	99	
8	232	19		477	96		895	112		549	98	
9	213	8	6	400	59	13	879	106	41	563	69	50
10	199	14	7	336	47	14	791	97	42	550	71	32
11	178	8	7	289	24	15	699	58	43	544	63	34
12	163	6	12	258	20	12	622	46	22	505	56	32
13	145	5	12	232	12	8	574	34	31	463	37	36
14	128	2	11	217	7	15	521	32	22	424	30	42
15	115	5	12	197	7	10	474	18	37	384	26	30
16	98	1	13	185	5	13	426	15	36	346	16	36
17	84	1	14	168	3	15	380	4	50	309	12	37
18	69	0	9	151	4	10	329	2	52	264	5	46
19	60	1	59	137	0	137	279	6	273	215	6	209

TABLE 2

Calculation of Life Table Probabilities of Having an ith Birth by Year of Marriage

Year of Marriage (x)	First Birth			Second Births			Third Births			Fourth Births		
	l_x^0	d_x^0	q_x^0	l_x^1	d_x^1	q_x^1	l_x^2	d_x^2	q_x^2	l_x^3	d_x^3	q_x^3
0	100,000	26,034	.26034									
1	73,966	31,682	.42833	26,034	3,807	.09091						
2	42,284	12,485	.29526	53,909	15,431	.25655	3,807	696	.06039			
3	29,800	8,596	.28846	50,963	16,619	.30074	18,542	4,134	.15396	696	82	.02963
4	21,204	5,567	.26255	42,940	13,058	.28558	31,027	7,941	.21144	4,748	1,433	.16432
5	15,637	2,824	.18063	35,449	9,946	.26985	36,141	7,561	.18616	11,255	2,489	.16554
6	12,813	2,047	.15974	28,327	7,737	.26360	38,525	7,260	.17126	16,327	3,792	.18999
7	10,766	1,269	.11787	22,637	4,380	.18821	39,002	6,683	.16223	19,795	4,043	.17476
8	9,497	778	.08190	19,526	3,930	.19733	36,699	4,592	.11877	22,435	4,006	.16198
9	8,719	332	.03809	16,374	2,455	.14843	36,037	4,252	.11410	23,021	2,936	.11675
10	8,387	601	.07161	14,251	2,035	.13988	34,240	4,186	.12224	24,337	3,203	.12118
11	7,786	357	.04584	12,817	1,092	.08406	32,089	2,745	.08412	25,320	3,025	.11331
12	7,429	284	.03822	12,082	959	.07843	30,436	2,290	.07407	25,040	2,864	.10938
13	7,145	257	.03597	11,407	601	.05206	29,105	1,771	.06023	24,466	2,030	.08008
14	6,888	112	.01633	11,063	370	.03325	27,935	1,752	.06232	24,207	1,796	.07160
15	6,776	311	.04587	10,805	394	.03599	26,553	1,049	.03922	24,163	1,698	.06878
16	6,465	71	.01093	10,722	300	.02793	25,898	952	.03654	23,514	1,141	.04767
17	6,394	83	.01299	10,493	196	.01863	25,246	285	.01122	23,325	963	.04102
18	6,311	0	.0	10,380	803	.07734	25,157	168	.00656	22,647	470	.02066
19	6,311	207	.03278	9,577	0	.0	25,792	372	.01444	22,326	1,190	.05286
TOTAL		93,897			84,113			58,689			37,161	

l_x^i = number of women of the cohort who reach parity i at the completion of x years after marriage

d_x^i = number of women who move from parity i to parity i + 1 during the xth year after marriage

q_x^i = the conditional probability of moving from parity i to parity i + 1 during the xth year after marriage

TABLE 3

Observed Timing of Transition from One Parity to the Next: 3,851 Selected White Women

Age	Women Eligible to Marry	Women Who Marry	Parity 0			Parity 1			Parity 2			Parity 3		
			Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date	Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date	Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date	Remaining at this Parity	Moved to Next Parity	At this Parity on Survey Date
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
13	3,851	8	0	1										
14	3,843	34	7	5		1	0							
15	3,809	91	36	28		6	2	1						
16	3,718	232	99	86	5	31	8	1	2					
17	3,486	383	240	185	8	108	32	7	10	2	1			
18	3,103	525	430	303	20	254	92	11	39	15	5	2		
19	2,578	524	632	381	26	454	148	28	111	29	6	17	2	4
20	2,054	518	749	380	31	659	188	38	224	71	13	40	15	3
21	1,536	445	856	376	31	813	245	23	328	83	18	93	22	13
22	1,091	301	894	337	44	921	269	44	472	111	29	141	42	12
23	790	206	814	277	23	945	272	42	601	120	36	198	60	14
24	584	157	720	229	21	908	251	31	717	156	52	244	53	19
25	427	96	627	195	17	855	225	35	760	154	42	328	62	27
26	331	89	511	146	15	790	222	22	789	146	30	393	70	31
27	242	51	439	100	11	692	194	22	835	140	35	438	82	30
28	191	46	379	79	12	576	128	16	854	128	39	466	74	28
29	145	32	334	60	7	511	94	18	815	120	41	492	72	21
30	113	24	299	49	4	459	104	13	748	90	32	519	62	31
31	89	26	270	41	8	391	71	13	730	68	42	516	65	39
32	63	17	247	37	8	348	51	13	691	75	42	480	45	36
33	46	14	219	18	16	321	49	13	625	45	18	474	48	36
34	32	8	199	22	10	277	24	12	611	41	30	435	24	32
35	24	3	175	19	7	263	20	12	564	39	26	420	29	31
36	21	4	152	9	8	250	23	12	519	25	36	399	28	36
37	17	3	139	6	10	224	10	8	481	17	40	360	20	27
38	14	5	126	3	10	212	11	14	434	16	30	330	14	26
39	9	2	118	4	9	190	3	13	399	6	32	306	14	24
40	7	1	107	3	6	178	2	14	364	6	43	274	6	36
41	6	1	99	1	7	165	3	25	317	6	33	238	4	41
42	5	2	92	1	10	138	4	13	281	3	41	199	3	25
43	3	1	83	1	9	122	0	15	241	1	38	174	4	28
44	2	1	74	0	10	108	2	16	202	0	36	143	0	35
45	1	0	65	1	4	90	0	12	168	1	34	108	0	22

TABLE 4

Calculation of Life Table Probabilities of Having an ith Birth by Age

Age	Marriage			First Birth			Second Birth			Third Birth			Fourth Birth		
	l_0	d_0	q_0	l_1	d_1	q_1	l_2	d_2	q_2	l_3	d_3	q_3	l_4	d_4	q_4
13	100,000	208	.00208												
14	99,792	883	.00885	-208	135	.20833									
15	98,909	2,367	.02393	956	735	.34355	135	51	.10256						
16	96,542	6,024	.06240	2,588	2,266	.40471	819	212	.10884	51	0	.00000			
17	90,518	9,961	.10987	6,346	4,901	.43274	2,873	865	.16234	263	55	.07843	0	0	.00000
18	80,557	13,629	.16919	11,406	8,089	.44396	6,909	2,519	.23000	1,073	424	.18181	55	0	.00000
19	66,928	13,630	.20365	16,946	10,276	.43246	12,479	4,135	.23473	3,168	834	.15934	479	61	.06780
20	53,298	13,441	.25219	20,300	10,345	.38287	18,620	5,389	.22650	6,469	2,089	.22793	1,252	466	.20270
21	39,857	12,528	.28971	23,396	10,491	.35371	23,576	7,136	.24760	9,769	2,507	.18799	2,875	710	.17187
22	27,329	7,540	.27590	25,433	9,625	.32960	26,931	8,005	.25219	14,398	3,450	.18750	4,672	1,410	.22047
23	19,789	5,160	.26076	23,348	7,122	.27467	28,551	8,221	.25600	18,953	3,849	.16689	6,712	2,064	.23904
24	14,629	3,944	.26963	21,386	6,788	.29061	27,452	7,688	.24925	23,325	5,191	.19105	8,497	1,881	.16960
25	10,685	2,402	.22482	18,542	5,776	.29257	26,552	7,084	.24064	25,822	5,310	.18085	11,807	2,290	.15836
26	8,283	2,227	.26888	15,168	4,338	.26642	25,244	7,143	.26056	27,596	5,142	.16497	14,827	2,703	.15538
27	6,056	1,304	.21074	13,057	2,989	.21806	22,439	6,351	.26538	29,597	5,017	.15309	17,266	3,289	.16632
28	4,752	1,144	.24085	11,372	2,383	.19949	19,077	4,271	.21070	30,931	4,711	.14246	18,994	3,062	.14341
29	3,608	796	.22069	10,133	1,824	.17316	17,189	3,198	.17669	30,491	4,576	.14260	20,643	3,049	.13296
30	2,812	597	.21239	9,105	1,491	.15857	15,815	3,618	.21849	29,113	3,550	.11479	22,170	2,707	.11303
31	2,215	647	.29213	8,211	1,254	.14695	13,688	2,509	.17530	29,181	2,780	.09133	23,013	2,992	.12260
32	1,568	423	.26984	7,604	1,150	.14711	12,433	1,843	.14169	28,910	3,217	.10783	22,801	2,199	.09009
33	1,145	348	.30435	6,877	582	.08257	11,740	1,822	.15147	27,536	1,998	.07025	23,819	2,489	.10031
34	797	199	.25000	6,643	749	.11111	10,500	925	.08510	27,360	1,876	.06743	23,328	1,325	.03460
35	598	75	.12500	6,093	673	.10983	10,324	822	.07707	26,409	1,859	.06933	23,879	1,697	.06839
36	523	100	.19047	5,495	333	.06000	10,175	926	.09255	25,372	1,260	.04878	23,717	1,732	.07115
37	423	75	.17647	5,262	316	.04428	9,582	437	.04484	25,038	921	.03648	23,245	1,355	.05714
38	348	124	.35714	5,021	122	.02429	9,461	507	.05326	24,554	940	.03769	24,165	1,061	.04307
39	224	50	.22222	5,023	176	.03493	9,076	148	.01617	24,121	377	.01550	24,044	1,142	.04713
40	174	25	.14286	4,897	141	.02870	9,104	106	.01159	23,892	418	.01746	23,279	544	.02316
41	149	25	.16667	4,781	50	.01041	9,139	179	.01953	23,580	470	.01986	23,153	367	.01568
42	124	50	.40000	4,756	54	.01136	9,010	274	.03030	23,289	338	.01442	23,256	335	.01428
43	74	25	.33333	4,752	61	.01282	8,790	0	.00000	23,225	105	.00450	23,253	582	.02496
44	49	25	.50000	4,716	0	.00000	8,851	173	.01960	23,120	0	.00000	22,776	0	.00000
45	24	24	.00000	4,741	75	.01587	8,678	0	.00000	23,293	154	.00662	22,776	0	.00000

TABLE 5

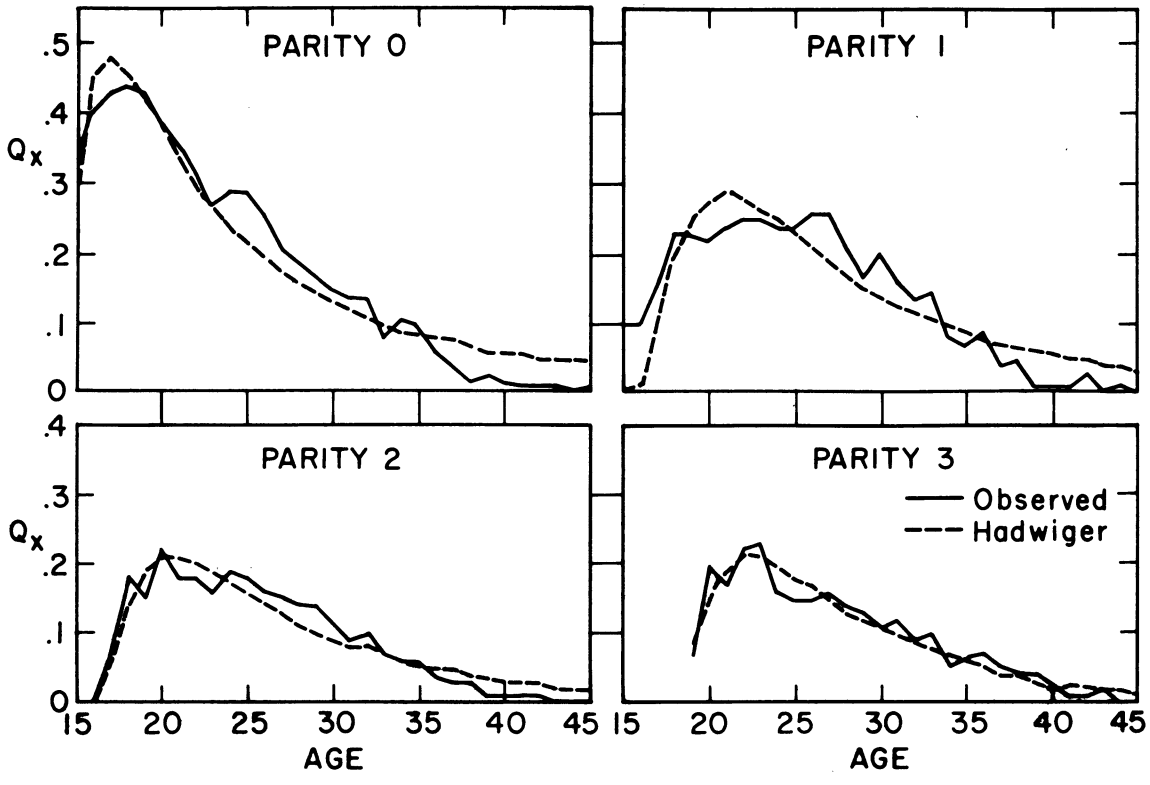
Observed Timing of Transition from One Parity to the Next: Grouped Data

Age	Women Who Marry	Parity 0			Parity 1			Parity 2			Parity 3		
		Remaining at This Parity	Moved to Next Parity	At This Parity on Survey Date	Remaining at This Parity	Moved to Next Parity	At This Parity on Survey Date	Remaining at This Parity	Moved to Next Parity	At This Parity on Survey Date	Remaining at This Parity	Moved to Next Parity	At This Parity on Survey Date
15-19	1755	36	983	59	6	282	48	0	46	12	0	2	4
20-24	1627	749	1,599	150	659	1,225	178	224	541	148	40	192	61
25-29	314	627	580	62	855	863	113	760	688	187	328	360	137
30-34	89	299	167	46	459	299	64	748	319	164	519	244	174
35-39	17	175	41	44	263	67	59	564	103	164	420	105	144
40-44	6	107	6	42	178	11	83	364	16	191	274	17	165

TABLE 6

Calculation of Life Table Probabilities from Grouped Data

Age	Parity 0		Parity 1		Parity 2		Parity 3	
	Average Years Spend in Parity Before Having Birth	q_x^i	Average Years Spend in Parity Before Having Birth	q_x^i	Average Years Spend in Parity Before Having Birth	q_x^i	Average Years Spend in Parity Before Having Birth	q_x^i
15-19	.3679	.9634	.3122	.9761	.2333	.7694	.1821	.3125
20-24	.6181	.9520	.5502	.8450	.4750	.7392	.5089	.6741
25-29	.6089	.7373	.6162	.7531	.5825	.5073	.5250	.5800
30-34	.5989	.5071	.5922	.5686	.6217	.3701	.5759	.3963
35-39	.5118	.2536	.6756	.2565	.6373	.1963	.6456	.2533
40-44	.5000	.0670	.7000	.0772	.5545	.0583	.7125	.0838



A STOCHASTIC PROCESS MODEL OF WORK FORCE HISTORY

P. Krishnan
University of Alberta

1 INTRODUCTION

Working life table is an accepted tool in demography. But this conventional technique has been found to be inappropriate for use with female populations of the developed societies in view of the bimodality of their labor force participation. Garfinkle (1968) has developed a procedure to take care of this feature of the modern female's work force participation. Since this requires refined data, the Garfinkle methodology cannot be put to use in most instances. Terry and Sly (1972) have adapted the working life table technique to get around the problem of bimodality by dividing the stationary work population into three components of those who (a) work continuously (b) work temporarily and (c) are temporarily out of work, and doing separate analysis of each of the components. These may be considered as ad hoc solutions to the bimodality problem. There are other problems as well which beset the labor force analysis. Some of these are pointed out elsewhere in the paper. These problems also require solutions for a better and meaningful characterization of the work force history.

The different sectors of the modern society are intricately interdependent. A change in one of them has immediate ramifications for others. This particularly applies to employment. The organized labor in one sector can precipitate temporary unemployment in others, when it votes to resort to strike action. Also the employment market is a highly competitive one. The supply of young inexperienced, highly qualified and sometimes overqualified people, racial and sex prejudices, mechanization, etc. result in some being hired, some early retired, and some fired.

In the developed societies, age 65 is considered as the age of retirement. In the developing nations, except for civil servants, there is no such concept of a retirement age. Persons have to work in order to survive. In some developed societies, some small segment of retired persons enter the labor force for personal and/or economic reasons. Others leave employment and later join the labor force for various other reasons (eg. to join a spouse who is transferred, or working elsewhere; join the graduate or technical school for higher education). In the actuarial type of methodology that is being employed in the construction of working life tables, the above discussed finer elements of labor force participation cannot be taken cognizance of. So we propose that the work force history of a person be looked at from a different perspective.

It is clear from the above discussion that the labor force history of a person can be characterized as a Markov Renewal process. Suitable modifications (eg. homogeniety assumption, use of mixing distributions) are needed to employ the model for a nation, or a large collectivity of people.

2 MARKOV RENEWAL PROCESS APPROXIMATION OF WORK FORCE HISTORY

At any point of time in one's life, one would be in one of the following states.

S_0 -Not in the labor force by not having entered it
S_1 -Employed
S_2 -Unemployed for involuntary reasons
S_3 -Unemployed for voluntary reasons (eg. to have a baby)
S_4 -Retired

S_2 and S_3 are further divisible if detailed information is available. To make the state space complete, we introduce S_5 the death state.

The length of stay of a person in state S_i before moving to S_j is a random variable with a distribution function $F_{ij}(t)$. The transition from S_i to S_j, in the appropriate unit of time, is governed by the elements of a transition probability matrix (P_{ij}). A typical labor force history is shown in Fig. 1. The model suggested here is more comprehensive than the ones in Hoem (1976, 1977) and Hoem and Fong (1976). Hoem and Fong have the age factor brought into the model directly. That can be accomplished here also by dividing the state space on the basis of age.

Some remarks are in order now. Since the death state is an absorbing state, all the first passage times are infinite. But still demographically meaningful results can be developed. All the elements P_{i5} (i=0, 1, 2, 3, 4) are the mortality rates specific for the labor force status. If suitable information is at hand, this model can make use of the differential mortality by work force status. Obviously, for either sex a separate model needs to be constructed.

NOTATION

We use the accepted notation in developing the results

T_{ij*} -wait in S_i before direct transition to S_j
T_{ij} -first passage time from S_i to S_j
$F_{ij}(t)$ -distribution function of the wait T_{ij*}
μ_{ij*} -Mean of $F_{ij}(t)$
μ_i $-\sum_j P_{ij}\mu_{ij*}$
m_{ij} $-E(T_{ij})$

The mean first passage times from all states to the death state can be easily derived by employing the following result due to Barlow and Proschan

Theorem (Barlow-Proschan)

Let [P, F(t)] be an absorbing semi-Markov process with k (0, 1, 2,...k-1) absorbing states where P has the normalized form $P=\begin{bmatrix} I & 0 \\ R & Q \end{bmatrix}$.

Then the mean time to absorption, starting in state i $(i > k)$ is $\sum_j m_{ij}\mu_j$ where $(m_{ij}) = (I-Q)^{-1}$.

3 SOME DEMOGRAPHICALLY USEFUL RESULTS

These results are not of much interest to us. We derive some other useful results. Let r_{ij} be the expected time in state S_j before death given that the person started from the state S_i.

r_{ij} = E[time in S_j before death/start from S_i]

Then we have the following:

r_{11} =expected life time in employed status given that the person started from his/her first job

r_{12} =expected life time in unemployed state (after joining the work force) for involuntary reasons

r_{13} =expected life time in unemployed state (after joining the work force) for voluntary reasons

r_{14} =expected life time in retirement

$\sum_{i=1}^{4} r_{ij}$ is the total life time after joining the work force for the first time. Then

$r_{ij}/\sum_{i=1}^{4} r_{ij}$ is the fraction of one's labor force life spent in status S_i. These are of demographic interest.

Before we go on developing the results, a look at the transition matrix P is called for. For definitional reasons P should have the following form

$$P = \begin{array}{c} \begin{matrix} S_0 & S_1 & S_2 & S_3 & S_4 & S_5 \end{matrix} \\ \begin{bmatrix} P_{00} & P_{01} & 0 & 0 & 0 & P_{05} \\ 0 & P_{11} & P_{12} & P_{13} & P_{14} & P_{15} \\ 0 & P_{21} & P_{22} & 0 & 0 & P_{25} \\ 0 & P_{31} & 0 & P_{33} & 0 & P_{35} \\ 0 & P_{41} & 0 & 0 & P_{44} & P_{45} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Let Y_{ij} be the random length of time in S_j before death given that the person started in S_i. For a first-step analysis, we have the following mutually exclusive possibilities:

a) the person can move directly from S_i $(i \neq 5)$ to the death state

b) the person can move from S_i to S_j directly and then from S_j to the death state

c) the person can move from S_i to S_k $(k \neq j, i)$ first and then from S_k to S_j from where the person moves to the death state

3.1 Theorem:

$$r_{11} = \mu_1 \Big/ 1 - \sum_{i=2}^{4} P_{1i}\, P_{i1}$$

Proof:

Now

$$Y_{11} = \begin{cases} T_{11*} & \text{Probability } P_{11} \\ T_{12*}+Y_{21} & \text{Probability } P_{12} \\ T_{13*}+Y_{31} & \text{Probability } P_{13} \\ T_{14*}+Y_{41} & \text{Probability } P_{14} \\ T_{15*} & \text{Probability } P_{15} \end{cases}$$

Then

$$r_{11} = E(Y_{11}) = \mu_1 + \sum_{i=2}^{4} P_{1i} r_{i1} \qquad (1)$$

To evaluate (1), we require the expressions for r_{21}, r_{31} and r_{41}.

Now

$$Y_{21} = \begin{cases} Y_{11} & \text{Probability } P_{21} \\ 0 & \text{Probability } 1-P_{21} \end{cases}$$

as the person has to move state from S_2 to S_1 with probability P_{21}. Then

$$r_{21} = P_{21}\, r_{11} \qquad (2)$$

Similarly

$$r_{31} = P_{31}\, r_{11} \qquad (3)$$

$$r_{41} = P_{41}\, r_{11} \qquad (4)$$

Substituting for r_{21}, r_{31}, and r_{41} from (2), (3), and (4), we get the expressions for r_{11} from (1).

3.2 Corollaries:

(a) $r_{21} = P_{21}\, r_{11}$

(b) $r_{31} = P_{31}\, r_{11}$

(c) $r_{41} = P_{41}\, r_{11}$

3.3 Theorem:

$$r_{12} = \frac{\mu_1 + P_{12}\mu_2 - P_{11}\mu_{11*}}{1 - \sum_{i=2}^{4} P_{1i}\, P_{i1}}$$

Proof:

$$Y_{12} = \begin{cases} T_{12*}+Y_{22} & \text{Probability } P_{12} \\ T_{13*}+Y_{32} & \text{Probability } P_{13} \\ T_{14*}+Y_{42} & \text{Probability } P_{14} \\ T_{15*} & \text{Probability } P_{15} \end{cases}$$

Then

$$r_{12} = (\mu_1 - P_{11}\mu_{11*}) + \sum_{i=2}^{4} P_{1i}\, r_{i2}$$

We have to evaluate r_{22}, r_{32} and r_{42};

$$Y_{22} = \begin{cases} T_{21*}+Y_{12} & \text{Probability } P_{21} \\ T_{22*} & \text{Probability } P_{22} \\ T_{25*} & \text{Probability } P_{25} \end{cases}$$

$$E(T_{22}) = r_{22} = \mu_2 + P_{21}\, r_{12}$$

Similarly

$$Y_{32} = \begin{cases} Y_{12} & \text{Probability } P_{31} \\ 0 & \text{Probability} \end{cases}$$

Thus $r_{32} = P_{31}\, r_{12}$

Similarly $r_{42} = P_{41}\, r_{12}$

$$\therefore r_{12} = (\mu_1 - P_{11}\,\mu_{11*}) + P_{12}[\mu_2 + P_{21}\,r_{12}] + P_{13}\,P_{31}\,r_{12} + P_{14}\,P_{41}\,r_{12}$$

$$r_{12}\left[1 - \sum_{i=2}^{4} P_{1i}\,P_{i1}\right] = \mu_1 + P_{12}\,\mu_2 - P_{11}\mu_{11*}$$

$$r_{12} = \frac{\mu_1 + P_{12}\,\mu_2 - P_{11}\,\mu_{11*}}{1 - \sum_{i=2}^{4} P_{1i}\,P_{i1}}$$

3.5 Corollaries:

(a) $r_{32} = P_{31}\,r_{12}$

(b) $r_{42} = P_{41}\,r_{12}$

Similarly we have

3.6 Theorem:

$$r_{13} = \frac{\mu_1 + P_{13}\,\mu_3 - P_{11}\,\mu_{11*}}{1 - \sum_{i=2}^{4} P_{1i}\,P_{i1}}$$

3.7 Theorem:

$$r_{14} = \frac{\mu_1 + P_{14}\,\mu_4 - P_{11}\mu_{11*}}{1 - \sum_{i=2}^{4} P_{1i}\,P_{i1}}$$

3.8 General Remarks:

The results derived above do not consider the effects of age, sex, education, etc. on work force history. These variables can be easily incorporated into the model by increasing the state space on the bases of these characteristics. Each of our S_i (i = 0, 1, ...5) could be thought of as states for each of the age groups for either sex and the various educational categories. The transition matrix would have then a large number of zero enteries.

REFERENCES

Garfinkle, S. (1968), "Work life experience of married women", Paper presented at the annual meetings of the Population Association of America, Boston.

Hoem, J. M. (1976), "The statistical theory of demographic rates", Scand. J. Statist, 3, 169-185.

Hoem, J. M. (1977), "A Markov Chain model of working life tables", Scand. Actuarial J., 1977, 1-20.

Hoem, J. M. and M. S. Fong (1976), "A Markov Chain model of working life tables", Working Paper No. 2, Laboratory of Actuarial Mathematics, University of Copenhagen, Copenhagen.

Terry, G. B. and D. F. Sly (1972), "A methodological alternative for constructing female work life tables", paper presented at the annual meetings of the American Statistical Association, Montreal.

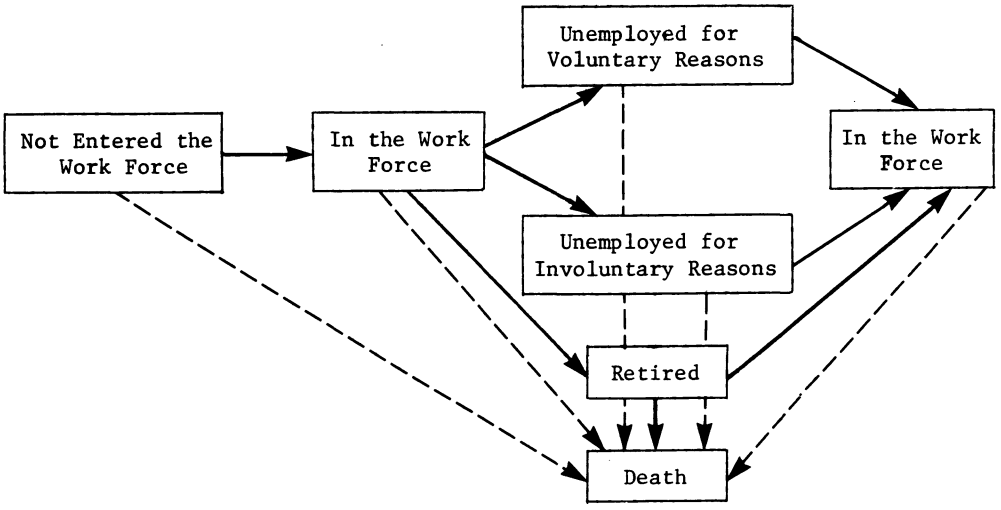
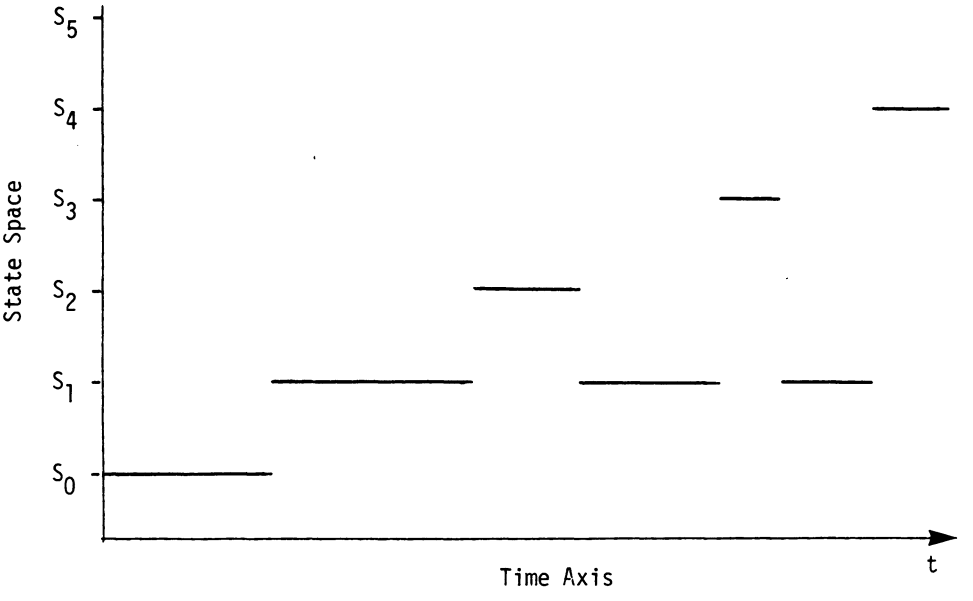


FIG 1A. A TYPICAL WORK FORCE HISTORY (Flow Chart)



Legend: S_0 - Not entered the work force

S_1 - In the work force

S_2 - Unemployed for involuntary reasons

S_3 - Unemployed for voluntary reasons

S_4 - Retired

S_5 - Death

FIG 1B. A TYPICAL WORK FORCE HISTORY

A MODIFIED CENSUS SURVIVAL RATIO METHOD FOR THE ESTIMATION OF INTERCENSAL NET INTERNAL MIGRATION

Alexander Mazurkewycz, Brandon University

As commonly applied, the forward census survival ratio method is defined as follows:

$$(1)\quad M_i = P_i^{(x+t,\,t)} - S_c P_i^{(x,o)}$$

where M_i is the net migration for the ith region, $P_i^{(x+t,\,t)}$ is the enumerated population in region i at age x + t at time t, $P_i^{(x,o)}$ is the enumerated population in region i aged x at time o, and S_c is the national survival ratio defined as

$$S_c = \frac{P_{\cdot}^{(x+t,\,t)}}{P_{\cdot}^{(x,\,o)}}$$

where P refers to summation over all i. The term $S_c P_i^{(x,o)}$ can be thought of as the "expected" population, expected in the sense that if there were no migration and the mortality conditions of the nation were evenly distributed, then this would be the "aged" population that we would expect.

The POBCSR method of Eldridge and Kim introduces a place of birth component into the method. Thus Eq. (1) becomes

$$(1a)\quad M_{ij} = P_{ij}^{(x+t,\,t)} - S_i P_{ij}^{(x,o)}$$

where M_{ij} refers to net migration into region j of the population born in i, P_{ij} refers to population residing at j and born in i, and

$$S_i = \frac{P_{i.}^{(x+t,\,t)}}{P_{i.}^{(x,\,o)}}$$

where $P_{i.}$ is population summed over all j, or in other words, total population born at i. It is clear that this technique is amenable to matrix manipulation so that we may now define the following:

$$P_{ij} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix}$$

and

$$M_{ij} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ m_{n1} & m_{n2} & & m_{nn} \end{bmatrix}$$

and Q a diagonal matrix with the main diagonal cells q_{ii} corresponding to the place-of-birth CSR's, S_i's, S_i's, where i = j. Equation (1a) can now be rewritten in matrix form:

$$(1b)\quad M_{ij} = P_{ij}^{(x+t,\,t)} - QP_{ij}^{(x,o)}$$

Let us further define a constant of proportionality C as follows:

$$(2)\quad C = \frac{P_{\cdot\cdot}^{(x+t,\,t)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{P_{j.}^{(x+t,\,t)}}{P_{j.}^{(x,o)}} P_{ij}^{(x,o)}}$$

where $P^{(x+t,\,t)}$ refers to the total population aged x + t at time t, and $P_{j.}$ refers to the total population born in region j irrespective of current place-of-residence. It is now possible to introduce a place-of-residence component into M_{ij} as follows:

$$(3)\quad M_{ij} = P_{ij}^{(x+t,t)} - \tfrac{1}{2}(QP_{ij}^{(x,o)} + CP_{ij}^{(x,o)}\,Q)$$

The essence of Eq. 3 is that net migration is the difference between the enumerated population at the time of the second census, and an expected population at that place, calculated on the basis of expected mortality. Thus, the expected population in Eq. 3 is equivalent to the term $\tfrac{1}{2}(QP_{ij}^{(x,o)} + CP_{ij}^{(x,o)}\,Q)$. amd consists of a place-of-birth census survival ratio, i.e. $QP_{ij}^{(x,o)}$ a place-of-residence survival ratio, i.e. $P_{ij}^{(x,o)}\,Q$.

Discussion of Method

It is important to remember, first of all, that Equation 3 is written in matrix form. Thus, the term $QP_{ij}^{(x,o)}$ has the effect of multiplying each row in $P_{ij}^{(x,o)}$ with the corresponding element in the diagonal matrix Q. This is the procedure advocated by Eldridge and Kim in the POBCSR method. The postmultiplication term $P_{ij}^{(x,o)}\,Q$ is equivalent to multiplying every column by the corresponding survival ratio in Q. The net effect of

these multiplications is that every cell in the $P_{ij}^{(x,o)}$ matrix gets a unique combination of survival ratios applied to it, i.e. a combined survival ratio of all those born at i and j. Since the CSR's are usually heavily weighted with these residing at place-of-birth, the combination of CSR's or Equation 3 can be justified as follows. The mortality experiences of a place-of-birth cohort may be combined with the mortality experiences of these born at place-of-residence because the place-of-residence CSR is heavily influenced by the population still residing at place-of-birth. The place-of-birth CSR is similarly weighted with those still residing at place-of-birth. Thus the method of Eq. 3 is a mean of the CSR's at place-of-birth and place-of-residence. This combination of mortality experiences should reflect actual mortality somewhat better than merely taking a place-of-birth CSR and applying it across a cohort.

It may appear initially that a more appropriate combination of survival ratios might include place-of-residence by using some formulation such as this: $S_j = P_{.j}^{(x+t,\,t)}/P_{.j}^{(x,o)}$. Although such a formulation might increase the mathematical elegance of the model, it does not help in interpreting any estimates since $P_{.j}$ does not consist of any identifiable cohort. Individuals may move into or out of a particular place at any time, therefore it makes little sense to construct a survival ratio for a group that is not closed in any way. Such a survival ratio already would have migration confounded within it, and there is no meaningful way of extracting the migration component from the mortality component. As a result, the constant of proportionality C must be introduced into the method.

The C coefficient is not as cumbersome and difficult to calculate as Eq. 2 may imply. It is merely the quotient of the total enumerated population at the second census divided by the overall total of all the cells of the matrix resulting from post-multiplication of the $P_{ij}^{(x,o)}$ matrix by the matrix operator Q. It is constant for every cohort, that is, for every $P_{ij}^{(x,o)}$ matrix.

A loss in elegance occurs when the POBCSR method is modified into the method of Eq. 3, or Combined CSR method for short; principally, no longer is the total expected population of a given place-of-birth equal to the total enumerated population of a place-of-birth. This is one of the strengths of the POBCSR method. The row totals Computed by the combined CSR method are off by a factor equal to (1 - C). However, this error is not greater than that incurred when the regular CSR method is employed, and most likely is less. Furthermore, this "error" is distributed throughout the population matrix in proportion to the size of each cell. The "error" involving each cell is not only contingent on the value of C, but is also dependent on the overall size of the P_{ij} matrix, and the relative frequency in each cell.

One further point regarding the constant of proportionality C requires elaboration. The closer that the value of C is to 1, the less "error" is involved in the migration estimates. This is largely a function of the number of cells in the overall population matrix P_{ij}, as well as the distribution of population in the matrix, and number of zero cells. Population matrices with disperse populations will produce C values that are quite close to 1. Because the value of C for a cohort is easily derived, it is apparent that a quick check is available to the researcher. If computed values of C deviate considerably from 1, say, by more than .05 or so, then an alternate method may be called for, perhaps the POBCSR method. If the C values lie close to 1, then some degree of confidence may be placed in migration estimates derived from the combined CSR method.

PROJECTION OF DIRECT FARM LABORER DISPLACEMENT FROM GEOTHERMAL DEVELOPMENT, IMPERIAL COUNTY, CALIFORNIA

James B. Pick, Tae Hwan Jung, and Edgar W. Butler
University of California, Riverside

ABSTRACT

Reduction in the farm laborer segment of the Imperial County labor force was projected based on losses in agricultural land directly caused by geothermal power plants and wells. A 100 MW power plant and well siting area was assumed to consume 650 acres of land. The proportions of land used in the well siting area by well pads, pipelines, access roads, possible subsidence, etc.--termed interstitial land reduction--were assumed at the levels of 5%, 10%, and 35%. Three scenarios of future power plant capacity in agricultural county areas were assumed. Ratios of farm laborers to land area, based on studies of Johnson (1977) and Sheehan (1976), were then used to project geothermal farm laborer displacement. For 35% interstitial land reduction, $4000 farm worker income, and the medium power plant scenario, the displacement is projected for year 2020 as only 1.96% of the 1970 farm laborer category.

INTRODUCTION

Geothermal energy resources exist as steam, hot water, and hot dry rock along tectonic plate fault lines in many parts of the world, including Sonoma County, north of San Francisco, and Imperial County, adjacent to Mexico in southeastern California. As part of a multidisciplinary project funded by NSF/ERDA, the farm labor impact of land consumption by geothermal development was investigated, and the results are the object of this paper. Other population and labor force aspects of this prospective energy development process have been detailed in previous reports (Pick et al., 1976; Pick, Jung, and Butler, 1977; Lofting, 1977; Rose, 1977), and a summary of the entire multidisciplinary project is available (Dry-Lands Research Institute, 1977).

Imperial County is a dry former desert, which due to irrigation diversions from the Colorado River beginning in 1904 has become one of the most fertile agricultural regions in the United States, producing about $1/2 billion of crops in 1976 from about 500,000 fertile acres in the central valley part. Because the central part of the county lies above a tectonic fault line known as the Salton Trough, there are large deposits of geothermal energy in the form of hot water located under the Imperial Valley at depths of 5-10,000 feet. One estimate of the recoverable energy capacity from these deposits is 10,000 MW over a 30-year lifetime (Biehler and Lee, 1976).

Geothermal development consists of the exploration and drilling of wells (somewhat equivalent to oil drilling) down to the depths of the hot water, transport of the hot water to the surface, and utilization of the hot water by flashing it to steam, to turn turbines in a power plant and generate electricity. Alternatively, the hot water can be used directly for house warming, air conditioning, industrial plant processes, etc.--uses referred to as non-electric.

With such a complicated energy source, there are many pathways that a geothermal development process can take, depending on such factors as total amount of recoverable energy, land ownership, permitting and regulatory processes, drilling costs, community and extra-local leadership, energy consumer market area, etc. It is impossible to project all such unknowns ahead of time, in part because there is only one U.S. geothermal field in active production--the Sonoma County steam resource with about 500 MW of installed electrical generating capacity. Hence projections of different types for Imperial County can only be performed with simplifying assumptions. County population projections were done, based on differing assumptions of buildup in geothermal capacity (Pick et al., 1976). These in turn were used to project county interindustry interactions (Lofting, 1977) and county revenues and taxes (Rose, 1977).

FARM LABOR FORCE REDUCTION BASED ON LAND AREA ANALYSIS

In Table 1 are presented the aggregated employment categories in Imperial County for the last three U.S. Censuses of Population. As expected for an agricultural county, the farming category is greatly enlarged relative to the U.S. as a whole. The 21% reduction in total percentage of the farm laborer category between 1960 and 1970 is exaggerated because of the presence of 4700-8000 border commuters, mostly farm laborers, who live in Mexicali, directly across the border, and commute to work daily in Imperial County. Such persons are not counted by the U.S. Census, since the Census counts persons based on residence (not workplace) in the U.S. (U.S. Senate, 1971). This group of 1970 commuting workers were mostly residents of the County in 1960, prior to the end of the Bracero Program (Samora, 1971).

The addition of the average of 6350 male commuting farm workers to the 1970 U.S. Census employment distribution (see Table 1) gives a 1970 farm worker fraction (40.4) quite similar to that in 1960 (36.9), and a 1970 total of 9537 farm laborers. Such a large proportion of county employment in this category warrants the special projections of the present paper. It is important to note that this total is also affected seasonally by harvesting cycles. Data in the present analysis are based on the Census date of April 15, even though maximal county employment due to crop cycles is in January.

The reduction from geothermal development of the farm laborer category in the Imperial County labor force was estimated based on prior studies of geothermal capacity (Davis, 1976), crop acreage (Johnson et al., 1977) and power plant impact (Sheehan, 1976; Rose, 1977). This analysis is based on the following schema for land reduction at one power plant site.

It is assumed that a 100 MW power plant installation will consume 10 acres (i.e., remove 10 acres from agricultural use by either direct or indirect effects) for the immediate area that the central power plant is sited on and the right-of-way for the central power plant. It is assumed that for the 100 MW capacity, there are 20 production wells and 12 reinjection wells. Each well is assumed to be spaced over 20 acres. The total well-spacing is thus assumed over 640 acres (slant drilling will be discussed below). The key question is then how much of the 640 acres is consumed either by well pads, pipelines, right-of-way, subsidence[1] problems (this will likely be small due to an assumed strong public policy against subsidence), and other environmental and agricultural causes. There are so many regulatory, agricultural, and geologic unknowns in the above causes that the present analysis simplified matters by assuming three possible percentages of land reduction for the well-spacing area (i.e., the 640 acres/100 MW capacity): (1) 5%, (2) 10%, and (3) 35%. These reductions are henceforth referred to as interstitial land reductions (abbreviated as i.l.r.).

For the case of slant drilling,[2] land is still consumed by centralized well pads, pipelines and right-of-way (albeit less), subsidence, and environmental-agricultural causes. Thus for slant drilling one should choose a smaller percentage reduction based on the partial land-consumption benefits of this method. Nevertheless, some land will still be consumed.

CALCULATIONS

The calculations of farm labor reduction are given in Tables 2 and 3. Two types of crop coverage are assumed for displaced KGRA[3]--an average 100 acres and all field crops. Table 2 presents the computations for an average 100 acres of crops as defined in a separate report by Sheehan (1976) and Rose (1977). For each KGRA a manpower reduction (in man-year units) is assumed based on the above report. Also, a scenario of power plant capacity is assumed for each KGRA based on the medium estimate of Davis (1976). The KGRA's are then summed in the right-hand columns to give total farm labor displaced by interstitial land reduction, farm income, KGRA, and year for field crop areas.

RESULTS

The general county results are given in Table 4. It is seen that the maximal reduction in labor force (year 2020, 35% i.l.r., average crops, $4,000 income) is 187 laborers, assuming the fixed agricultural mechanization and other trends. Based on a present labor force (including 6350 border commuters) of 29,829, this is a reduction of only .62%. Since it is likely that the county will increase in population (Pick et al., 1976), this percentage may be reduced by 50% based on year 2020 populations. The 5% i.l.r. for the other above assumptions unchanged yields a farm labor reduction of only 27 laborers, or .09% based on 1970 population. For the case of all field crops, the figures are roughly 75% less than for average crops. Based on a 1970 farm laborer category (including border commuters) of 9537, the 35% i.l.r. and 5% i.l.r. reductions are 1.96% and .28% respectively.

ACKNOWLEDGMENTS

The authors thank Stahrl Edmunds and Mike Pasqualetti for helpful advice and comments. Funding was provided by NSF/ERDA Grant AER 75-08793.

NOTES

[1] Subsidence is the lowering of land levels in a geothermal production area due to withdrawal of geothermal fluids.

[2] Slant drilling is a drilling process where the angle of the drill to the land surface is significantly different than 90 degrees.

[3] A Known Geothermal Resource Area (KGRA) is a federal designation of a land area of geothermal deposits, based both on geologic and potential commercial characteristics.

REFERENCES

Biehler, Shawn and Tien Lee. 1976. Final Report on a Resource Assessment of the Imperial Valley. Riverside, Dry-lands Research Institute, University of California.

Davis, C. 1976. Preliminary Scenario, Imperial Valley Region, State of California, unpublished report. Pasadena, Jet Propulsion Laboratory.

Dry-lands Research Institute. 1977. Summary of Final Reports on Geothermal Energy Development in Imperial County. Riverside, University of California.

Johnson, C. 1977. Effects of Geothermal Development on the Agricultural Resources of the Imperial Valley. Riverside, Dry-lands Research Institute, University of California.

Lofting, E. M. 1977. A Multisector Analysis of the Imperial County Economy. Riverside, Dry-lands Research Institute, University of California.

Pick, J. B., T. H. Jung, and E. W. Butler. 1977. Regional Employment Implications for Geothermal Energy Development, Imperial County, California. Proceedings of the Los Angeles Council of Engineers and Scientists, 3:159-167.

Pick, J. B., C. Starnes, T. H. Jung, and E. W. Butler. 1976. Population-Economic Data Analysis Relative to Geothermal Fields, Imperial County, California. Proceedings of the American Statistical Association, Social Statistics Section, II:666-672.

Rose, A. 1977. The Economic Impact of Geothermal Energy Development. Riverside, Dry-lands Research Institute, University of California.

Samora, Julian. 1971. Los Mojados: The Wetback Story. Indianapolis, University of Notre Dame Press.

Sheehan, Mike. 1976. Report of the Estimated Impact of Locating a One Hundred Acre Geothermal Facility on Various Lands. Preliminary report, NSF-Imperial County Geothermal Project, August.

U.S. Bureau of the Census, 1973. U.S. Census of Population, 1970. General Population Characteristics. Final Report PC(1)-B. Washington, D.C., U.S. Government Printing Office.

U.S. Senate, 91st Congress. 1971. Migrant and Seasonal Farm Worker Powerlessness. Hearings on Border Commuter Problem. Washington, D.C., U. S. Government Printing Office.

TABLE 1. MALE EMPLOYMENT CATEGORIES FOR IMPERIAL COUNTY 1950-1970

	1950	1960	1970	1970*
Total Male Labor Force	18599	21613	15397	21747
Employment Category:				
Aggregated Categories:				
Farm (Including Farm Managers)	40.1	43.2	20.7	43.8
Farmers and Farm Managers	9.2	6.3	4.8	3.4
Farm Laborers and Foremen	30.9	36.9	15.9	40.4
Clerical and Sales	7.4	6.9	10.4	7.4
Professionals and Managers	13.9	14.6	22.2	15.7
Craftsmen and Operatives	26.0	23.5	31.9	22.6
Other	12.6	11.8	14.8	10.4

*border commuters included (see text)

Source: U. S. Bureau of the Census, 1950-70

TABLE 2. FARM LABOR FORCE DISPLACEMENT BY KGRA, PLANT CAPACITY, FARM WORKER INCOME, AND LAND REDUCTION FOR AVERAGE 100 ACRES

		KGRA											
		Salton Sea			Brawley			Heber			Total		
	Year	Capacity[1]	FLD5	FLD4	Capacity	FLD5	FLD4	Capacity	FLD5	FLD4	Capacity	FLD5	FLD4
	1980	90	.51	.64	35	.18	.22	35	.66	.83	180	1.35	1.69
	1990	250	1.42	1.77	100	.51	.64	100	1.89	2.36	500	4.77	5.97
5% i.l.r.	2000	1,250	7.16	8.95	500	2.55	3.19	500	9.47	11.84	2,500	19.18	23.97
	2010	1,750	10.02	12.52	700	3.57	4.46	700	13.25	16.56	3,500	26.84	33.55
	2020	1,750	10.02	12.52	700	3.57	4.46	700	13.25	16.56	3,500	26.84	33.55
	1980	90	.91	1.14	35	.31	.39	35	1.17	1.46	180	2.39	2.99
	1990	250	2.52	3.15	100	.90	1.12	100	3.33	4.17	500	6.75	8.44
10% i.l.r.	2000	1,250	12.62	15.77	500	4.49	5.62	500	16.68	20.85	2,500	33.79	42.24
	2010	1,750	17.66	22.07	700	6.30	7.87	700	23.35	29.19	3,500	47.31	59.14
	2020	1,750	17.66	22.07	700	6.30	7.87	700	23.35	29.19	3,500	47.31	59.14
	1980	90	2.87	3.59	35	.99	1.24	35	3.69	4.61	180	7.55	9.44
	1990	250	7.98	9.97	100	2.84	3.56	100	13.18	16.48	500	24.00	30.00
35% i.l.r.	2000	1,250	39.90	49.87	500	17.78	22.23	500	52.74	65.93	2,500	110.42	138.02
	2010	1,750	55.85	69.82	700	19.92	24.90	700	73.84	92.30	3,500	149.61	187.01
	2020	1,750	55.85	69.82	700	19.92	24.90	700	73.84	92.30	3,500	149.61	187.01

i.l.r. = interstitial land reduction
FLD5 = farm labor displacement (average annual income = 5,000)
FLD4 = farm labor displacement (average annual income = 4,000)
All capacity figures in MW

TABLE 3. FARM LABOR FORCE DISPLACEMENT BY KGRA, POWER PLANT CAPACITY, FARM WORKER INCOME, AND LAND REDUCTION PERCENTAGE FOR FIELD CROPS

		KGRA											
		Salton Sea			Brawley			Heber			Total		
	Year	Capacity	FLD5	FLD4	Capacity	FLD5	FLD4	Capacity	FLD5	FLD4	Capacity	FLD5	FLD4
	1980	90	.23	.29	35	.07	.08	35	.07	.08	180	.37	.46
	1990	250	.64	.80	100	.19	.24	100	.19	.24	500	1.02	1.27
5% i.l.r.	2000	1,250	3.20	4.00	500	.96	1.20	500	.97	1.21	2,500	5.13	6.41
	2010	1,750	4.48	5.61	700	1.34	1.67	700	1.36	1.70	3,500	7.18	8.97
	2020	1,750	4.48	5.61	700	1.34	1.67	700	1.36	1.70	3,500	7.18	8.97
	1980	90	.41	.51	35	.12	.15	35	.12	.15	180	.65	.81
	1990	250	1.13	1.41	100	.34	1.42	100	.34	.42	500	1.81	2.26
10% i.l.r.	2000	1,250	5.66	7.07	500	1.68	2.10	500	1.71	2.14	2,500	9.05	11.31
	2010	1,750	7.92	9.90	700	2.36	2.95	700	2.40	3.00	2,500	12.68	15.85
	2020	1,750	7.92	9.90	700	2.36	2.95	700	2.40	3.00	3,500	12.68	15.85
	1980	90	1.29	1.61	35	.37	.46	35	.38	.47	180	2.04	2.55
	1990	250	3.58	4.47	100	1.06	1.32	100	1.08	1.35	500	5.72	7.15
35% i.l.r.	2000	1,250	17.88	22.35	500	5.32	6.65	500	5.41	6.76	2,500	28.61	35.76
	2010	1,750	25.04	31.30	700	7.45	9.37	700	7.57	9.46	3,500	40.06	50.07
	2020	1,750	25.04	31.30	700	7.45	9.37	700	7.57	9.46	3,500	40.06	50.07

i.l.r. = interstitial land reduction
FLD5 = farm labor displacement (average annual income = $5,000 in 1970 dollars)
FLD4 = farm labor displacement (average annual income = $4,000 in 1970 dollars)
All capacity figures in MW

TABLE 4. REDUCTION IN THE NUMBER OF FARM LABORERS FROM GEOTHERMAL DEVELOPMENT--JPL SCENARIO

	1980	1990	2000	2010	2020
5% i.l.r., field, M, $5,000	.37	1.02	5.13	7.18	7.18
5% i.l.r., field, M, $4,000	.46	1.27	6.41	8.97	8.97
10% i.l.r., field, M, $5,000	.65	1.81	9.05	12.68	12.68
10% i.l.r.,field, M, $4,000	.81	2.26	11.31	15.85	15.85
35% i.l.r., field, M, $5,000	2.04	5.72	28.61	40.06	40.06
35% i.l.r., field, M, $4,000	2.55	7.15	35.76	50.07	50.07
5% i.l.r., aver., M, $5,000	1.35	4.77	19.18	26.84	26.84
5% i.l.r., aver., M, $4,000	1.69	5.97	23.97	33.55	33.55
10% i.l.r., aver., M, $5,000	2.39	6.75	33.79	47.31	47.31
10% i.l.r., aver., M, $4,000	2.99	8.44	42.24	59.14	59.14
35% i.l.r., aver., M, $5,000	7.55	24.00	110.42	149.61	149.61
35% i.l.r., aver., M, $4,000	9.44	30.00	138.02	187.01	187.01

i.l.r. = interstitial land reduction

field = all field crops

aver. = average crop distribution (Sheehan, 1976)

M = middle Cal Tech power plant capacity Scenario (Davis, 1976)

dollar values = income figures for farmworkers (1970 dollars)

DEMOGRAPHIC CORRELATES OF ATTITUDES TOWARDS BUSING AND SCHOOL INTEGRATION: DETROIT, 1976

John B. Casterline, University of Michigan, Population Studies Center

In the 1976 Detroit Area Study respondents were asked a series of questions on school busing and the racial integration of schools. These questions may be set in the context of nearly a decade of controversy in the Detroit metropolitan area over the busing of school children, in which, following nation-wide sentiment, opposition to busing has overwhelmed support. During the same period, surveys have found equally strong support, both in Detroit and nation-wide, for the notion of integrated schools. In fact, the intense opposition to busing observed in the U.S. during the past decade has arisen during a period in which racial integration has become accepted in most public realms of U.S. society, including the schools (Sheatsley, 1966; Campbell, 1971; Greeley, 1971). The strength of the opposition to busing, and the anomaly of opposition to busing in the midst of a general liberalization of racial attitudes in the U.S. (if the evidence from the national surveys is believed) stimulates a number of questions which this paper will address: (1) Are support and opposition to school busing located in different sub-groups of the population than the sub-groups supporting and opposing school integration? That is, given that school busing is much less popular than the concept of school integration, do we still observe the differences in support or opposition by major sub-groups of the population--race, age, education sub-groups--which have been found in studies of racial attitudes in general, and in attitudes towards school integration in particular? The bulk of this paper is an examination of these questions. (2) What implications for the career of school busing as a public issue are suggested by these findings on the social location of its support and opposition? This question is considered briefly at the end of the paper.

An extensive literature explores the relationship of demographic characteristics to racial attitudes (Allport, 1962; Schwartz, 1967; Campbell and Schuman, 1968; Campbell, 1971; Pettigrew, 1971; Schuman and Hatchett, 1974). A smaller number of studies focus on attitudes towards school integration in particular (Schwartz, 1967; Greeley and Sheatsley, 1971; Pettigrew, 1971; Giles, et. al., 1976). Only a few studies have examined attitudes towards busing (Crain, 1968; Rubin, 1972; Kelley, 1974). This literature is limited to considerations of White racial attitudes almost exclusively; the literature on Black racial attitudes is scanty (Marx, 1967; Caplan and Paige, 1968; Paige, 1969; Edwards, 1972; Schuman and Hatchett, 1974) and primarily concerned with the correlates of Black militancy.

A number of strong relationships have emerged from this previous work. First, the results of all surveys which included White and Black respondents have confirmed significant differences in attitudes by race, with Blacks as a group more approving of racial integration--in the schools and elsewhere.

Second, age of respondent has usually been found inversely related to liberal attitudes on racial issues among Whites (Sheatsley, 1966; Campbell and Schuman, 1968; Campbell, 1971; Pettigrew, 1973; Smith, 1976) and, among Blacks, inversely related to militancy and anti-White attitudes (Marx, 1967; Paige, 1970; Schuman and Hatchett, 1974).

Third, studies of White racial attitudes typically uncover a positive relationship of educational attainment of respondents with liberal racial attitudes (Schwartz, 1967; Campbell and Schuman, 1968; Campbell, 1971; Greeley and Sheatsley, 1971; Smith, 1976), while a few investigations of Black racial attitudes discover greater militancy among the most and the least educated (Caplan and Paige, 1968; Schuman and Hatchett, 1974).

Considered together, these previous studies suggest that both age and race, and education and race, may interact in influencing racial attitudes. Moreover, previous work on both White and Black racial attitudes points to an interaction of age and education (Campbell, 1971; Schuman and Hatchett, 1974); for both races, education seems to be more strongly related to the responses of younger respondents than to those of respondents over 40 years old.

These three demographic variables--race, age, and education--have received considerable attention in the racial attitude literature. In this study we introduce another variable for consideration which has not been included in most studies--having children in the public schools. Many of the emotions and issues involved in the school busing controversy would seem to be more salient for parents of children in the schools than for non-parents (for example: fears of racial conflict in the schools; anxiety among parents about sending children far away to school; concern about educational standards in the schools; concern among Blacks about unsympathetic administrators and teachers in predominantly White schools). We expect that having children in the public schools may directly affect responses to a question on school busing or may interact--with race or education or age--in influencing responses.

This paper examines the relationship of these four variables--race, education, age, and having children in the public schools--to two questions from the 1976 Detroit Area Study: one question asked whether the respondent approved or disapproved of school busing as a means to integrate the schools of Detroit; the second asked whether the respondent would object to sending his or her children to a school where more than half of the children are of the opposite race. The latter question we select from a series of items on the issue of school integration. (These two questions from the interview are given in Appendix A, attached to the tables. Appendix B shows the correlations between the four demographic variables). Our analysis proceeds with two main questions in mind: Do these demographic variables influence attitudes towards busing? Do these demographic variables

influence attitudes towards busing? Do these demographic variables influence similarly attitudes towards busing and attitudes towards having one's children in a school where more than half of the children are of the opposite race?

The 1976 Detroit Area Study sample is particularly useful for this analysis because 400 of the 1134 respondents were Black. (238 of these 400 were selected as a Black supplement.) This is far more Blacks than often obtained in major national studies and allows us to analyze multidimensional contingency tables which include race along with other demographic variables. The 1134 interviews were obtained from April to August 1976 from a probability sample of the entire Detroit SMSA; the overall response rate was 75.4%.

To simplify the log-linear analysis which follows, all the variables, with one exception, are dichotomized, after curvilinear relationships were searched for and not found. Skew in the distribution of responses to the busing question forced a trichotomizing of this variable into the categories "Approve", "Disapprove", and "Strongly Disapprove".

Before proceeding into the log-linear analysis, we may note in Tables 1 and 5 that only about 7% of the Whites in the sample approve of busing, while 50% of the Blacks approve. A higher percentage of Whites--33%--have no objection to sending their children to schools where more than half of the children are of the opposite race, but 84% of the Blacks do not object to this circumstance. These tables indicate large differences in responses by race and extreme opposition to busing on the part of Whites.

Following the methods of log-linear analysis which are now common in sociological research (Goodman, 1971; Goodman, 1972; Davis, 1974), we examine the multivariate contingency tables produced by the exhaustive classification of the four demographic variables against first, attitudes towards busing, and, later, attitudes towards sending one's children to schools where more than half of the children are of the opposite race. The methods of log-linear analysis are especially useful for the problem at hand because they allow the explicit testing of interactions among variable in the tables, as well as the testing of direct effects of the independent variables on the dependent variable. The log-linear models are fit to the multivariate tables by an iterative proportional fitting procedure as implemented by the ECTA computer program.

Table 6 displays the observed data for the five-way table of BUSING by race by children in the schools by age by education. Table 7 presents the models fit to these data, and Table 8 compares the fit of selected pairs of these models and tests the significance of specific terms. We note first that the differences between model 1 and models 2, 3, 4, and 5, taken singly, provide tests of the significance of the direct effects of education, age, children, and race on BUSING. Only race proves to have a significant effect (only model 5 stands as an improvement in fit over model 1). Likewise, the differences between model 6 and models 7, 8, 9, 10, 11, and 12, taken singly, provide tests of the significance of the interaction effects on attitudes towards busing of education and age, education and race, education and children, age and race, age and children, and race and children. Only the education-age interaction proves significant. Finally, the differences between model 13 and models 14, 15, 16 and 17 provide tests of the significance of four four-variable interaction terms. None tests as significant.

Summarizing the results of these tests: only race of respondent, and the interaction of education and age of respondent, affect responses to the busing question. Direct effects of education and age of respondent which we might have expected from previous racial attitude research do not appear; nor does the direct effect of having children in the public schools which we hypothesized. Interactions of race and education which we might also have expected do not test as significant. In regard to those terms which do test as significant space does not allow a full presentation of the effects of these terms on busing responses. Logit effects parameters were computed for the "best-fitting" model shown in Table 7, model 18 (EAB, RB, EACR). (The effects parameters are not shown here; author will supply them if requested.) These effects parameters indicate that the race-busing association is far more powerful than the education-age-busing interaction. Whites are much more disapproving of busing than Blacks, of course. And, while the education-age interaction is not completely clear, there is some indication that age has more effect among those with more education.

A glance at Tables 2 through 4 confirm for this Detroit sample what was discussed at the beginning of this paper: strong opposition to busing is combined with overwhelming support for the concept of racial integration of the schools (Table 2) and lack of objection to sending one's children to schools where a few or half of the children are of the opposite race (Tables 3 and 4). However, Table 5--attitude towards sending one's children to a school where more than half of the children are of the opposite race--shows again a strong division between White and Black respondents, with the majority of Whites opposed to sending their children to such schools. Yet--if we may for a moment consider current realities--school integration in the Detroit area (certainly in Detroit city proper) would require many White children to attend schools where they would be a minority. (At present, 81% of the students in Detroit city schools are Black.) White respondents appear to support school integration as an ideal but reject the circumstances which follow (e.g., school busing, minority status for some children in the schools) from any rapid implementation of that ideal.

We proceed to examine the responses to the question asking about sending one's children to a school where most children are of the opposite race; our primary purpose is to see whether the relationships of the responses to this question (which I shall term the "MIXING" question) with the independent variables are similar or differ from those for the responses to the busing question.

Table 9 displays the observed data for the five-way table of MIXING by race by children by age by education. Table 10 presents the models fitted; Table 11 compares the fit of selected pairs of these models. The differences between model 1 and models 2, 3, 4, and 5, taken singly, provide tests of the significance of the direct effects of education, age, children, and race on school mixing. Age and race prove to have significant effects. The three-variable interactions are tested by the differences between model 6 and models 7 through 12. Four interactions--education and age, education and race, age and race, and age and children--test as significant in their effects on responses to the MIXING question. The four-variable interactions are tested by the differences between model 13 and models 14 through 17. None of these test as significant at the .01 level, although education-race-children and education-age-children quite nearly do.

Summarizing the results of these tests: Both age of respondent and race of respondent directly affect responses to this item. Four interactions--including several (e.g. education and age, education and race, age and race) which previous investigations led us to expect to possess explanatory power--also tested as significant. Again space does not allow a full presentation of these relationships. Logit effects parameters were compiled for the "best-fitting" model shown in Table 12, Model 20 (EACM, ARCM, EACR). (The effects parameters are not shown here.) At least one relationship--age is inversely related to no objection to sending one's children to a school where more than half of the children are of the opposite race--runs counter to expectations based on the racial attitude literature, although perhaps not counter to common-sense expectations on this particular issue. On the other hand, race influences the responses as we would expect--Whites object much more frequently than Blacks--and the magnitude of the effects parameters indicate that race has more effect than any of the other variables considered. The nature of the interactions which test as significant in Table 11 is not best examined by looking at the effects parameters for the best-fitting model, Model 20. They do suggest, however, that age and race interact such that age has a greater effect on the responses of Blacks than Whites; education and age seem to interact such that education has more effect on those persons under 40 than those over 40.

We conclude our findings with a few summary comparisons of the results of the analysis of the busing and MIXING tables. In both cases race of respondent has the most powerful effect on the distribution of responses observed. When we proceed beyond that direct effect, we find that <u>different</u> models fit the two sets of data; the same relationships between the dependent and independent variables do <u>not</u> hold when the two sets of responses are examined. Furthermore, if we require a p value of .500 for a model to qualify as "best-fitting", we note that the analysis of the table for busing yields a best-fitting model with nothing more complex than a race-busing association and an education-age-busing interaction, whereas analysis of the table for MIXING yields a best-fitting model consisting of two interactions of MIXING with three other variables. These comparisons suggest that there is simply much more "action" among the independent variables and the dependent variable in the table of responses to the school MIXING question than there is among the independent variables and the dependent variable in the table of responses to the busing question. In the latter table, knowing the race of respondent gets us quite a long ways towards an accurate prediction of responses to the busing question.

The results of the above analysis emphasize what political and social observers following the busing controversy in Detroit and elsewhere have suggested: opposition to the busing of school children in order to integrate schools is spread throughout the socity. Demographic characteristics which past studies have found related to racial attitudes and attitudes towards school desegregation seem to hold no force here. Race is the sole characteristic which carries weight: Blacks generally approve of busing, while Whites oppose it. Education and age prove to have no direct influence; a relatively weak interaction of the two is suggested by the results, but it would be a mistake to emphasize its importance since the race effect is of much greater magnitude. Having children in the schools also proves to have no direct effect on attitudes towards busing, running against our instincts (but repeating Kelley's findings in his analysis of the 1972 NORC data: Kelley, 1974). And, with the exception of the education-age interaction just noted, none of the interactions which use of the log-linear analysis allowed us to test proves significant.

Examination of the multidimensional contingency table of the responses to a question on having one's children in a school where more than half of the children are of the opposite race, produced different results. Race remains of overwhelming importance, but other associations and interactions test as significant. Moreover, a good fit to the observed data requires a more complicated model than was required for the busing question.

Comparing the results of the analysis of the two data sets suggests that attitudes towards busing are distinctive by their absence of demographic correlates, beyond race of respondent. We speculate, nevertheless, that the comparison was not ideal because the question on school integration used in the comparison may be a weak measure of integration sentiments. If, instead, we compare our findings with those of others who have studied the demographic correlates of attitudes towards school integration (or, indeed, racial attitudes in general), then the anomaly of the <u>absence</u> of demographic correlates of busing attitudes stands out more strongly.

Approval and disapproval of school busing do not seem to be located socially the way other racial issues have been. Consequently, perhaps a different career should be expected for school busing as a public issue than has been observed for racial issues. In the case of many of these issues, gradual public acceptance of more liberal

policy has followed initial opposition to liberal policy (Greeley and Sheatsley, 1971). Usually the upper socioeconomic strata (especially the more educated) and the young have been the vanguard groups in the change. These data from Detroit, however, provide little evidence that the better-educated and the young are more approving of school busing. (Nor, it might be added, do we find any evidence that the young and better-educated are more willing to allow their children to attend schools where they would be in a minority.) There do not appear to be "vanguard groups"--certainly not among Whites--in Detroit.

ACKNOWLEDGMENT

This research was sponsored, in part, by a grant from the National Science Foundation, SOC76-00768, "Causes of Racial Residential Segregation in the Detroit Area." Helpful comments on this and an earlier draft were provided by Reynolds Farley, Robert Groves, and Paul Siegel.

APPENDIX A

Description of the Variables

BUSING: "The courts ordered busing for some Detroit school children this past winter. What do you think of busing as a way to integrate the schools of the city of Detroit? Do you strongly approve, approve, disapprove, or strongly disapprove of busing for integration in Detroit?"

In this analysis:
Category 1. "Approve": "Strongly approve" and "Approve" responses
Category 2. "Disapprove": "Disapprove" responses
Category 3. "Strongly disapprove": "Strongly disapprove" responses

MIXING: "Would you have any objection to having children of your own attend a school where more than half of the children are (OPPOSITE RACE OF R)?"

In this analysis:
Category 1: "No objection": "No objection" responses
Category 2: "Object": "Object" responses to this question plus respondents who objected to sending their children to a school where a few or half of the children are of the opposite race. (Respondents who objected to sending their children to a school where a few of the children are of the opposite race were not asked whether they objected to sending their children to a school where half or more than half are of the opposite race. Likewise, those who objected to sending their children to a school where half are of the opposite race were not asked whether they objected to sending them to a school where more than half are of the opposite race.)

RACE:
Category 1: White
Category 2: Black
"Other" category excluded from this analysis

CHILDREN: "Do you have any children attending public schools?"
Category 1: Yes
Category 2: No

AGE:
Category 1: Less than 40 years old. Respondents range from 18 to 39 years old.
Category 2: 40 years or older. Respondents range from 40 to 93 years old.

EDUCATION:
Category 1: 12 years of schooling or less. Respondents range from 0 to 12 years of schooling.
Category 2: More than 12 years of schooling. Respondents range from 12 to 17+ years of schooling.

APPENDIX B

Correlations (Gammas) between the Independent Variables

	RACE	CHILDREN	AGE	EDUCATION
RACE	--	-.06	-.08	-.24
CHILDREN		--	.41	.21
AGE			--	-.35
EDUCATION				--

REFERENCES

Allport, Gordon W.
1962 "Prejudice: Is it Societal or Personal?", Journal of Social Issues, Vol. 18, pp. 120-134.

Campbell, Angus
1971 White Attitudes Toward Black People. Ann Arbor: Institute for Social Research.

Campbell, Angus and Howard Schuman
1968 "Racial Attitudes in Fifteen American Cities", in Supplemental Studies for the National Advisory Commission on Civil Disorders. U.S. Government Printing Office.

Caplan, Nathan S. and Jeffery M. Paige
1968 "A Study of Ghetto Rioters", in Scientific American, Vol. 219, No. 2, pp. 15-21.

Crain, Robert L.
1968 The Politics of School Desegregation. Chicago: Aldine Publishing Company.

Davis, James A.
1974 "Hierarchical Models for Significance Tests in Multivariate Contingency Tables: An Exegesis of Goodman's Recent Papers", in H.L. Costner, ed., Sociological Methodology, 1973-1974, San Francisco: Jossey-Bass, Inc.

Edwards, Ozzie L.
1972 "Intergenerational Variation in Racial Attitudes", Sociology and Social Research, Vol. 57, No. 1, pp. 22-31.

Giles, Michael W., Douglas S. Gatlin, Everett F. Cataldo
1976 "Racial and Class Prejudice: Their Relative Effects on Protest against School Desegregation", American Sociological Review, Vol. 41, No. 2, pp. 280-288.

Goodman, L.A.
1971 "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications", Technometrics, Vol. 13, pp. 33-61.

1972 "A General Model for the Analysis of Surveys", American Journal of Sociology, Vol. 77, No. 6, pp. 1035-1086.

Greeley, Andrew M. and Paul B. Sheatsley
1971 "Attitudes toward Integration", Scientific American, Vol. 225, pp. 13-19.

Kelley, Jonathan
1974 "The Politics of School Busing", Public Opinion Quarterly, Spring, 1974, pp. 23-39.

Marx, Gary T.
1967 Protest and Prejudice. New York: Harper & Row.

Paige, Jeffery
1970 "Changing Patterns of Anti-White Attitudes Among Blacks", Journal of Social Issues, Vol. 26, No. 4, pp. 69-86.

Pettigrew, Thomas F.
1971 Racially Separate or Together? New York: McGraw-Hill.

Pettigrew, Thomas F.
1973 "Attitudes on Race and Housing: A Social-Psychological View", in Amos Hawley and Vincent Rock, eds., Segregation in Residential Areas. Washington, D.C.: National Academy of Sciences, pp. 21-84.

Rubin, Lillian B.
1972 Busing and Backlash. Berkeley: University of California Press.

Schuman, Howard and Shirley Hatchett
1974 Black Racial Attitudes: Trends and Complexities. Ann Arbor: Survey Research Center.

Schwartz, Mildred A.
1967 Trends in White Attitudes toward Negroes. Chicago: National Opinion Research Center.

Smith, A. Wade
1976 "Tolerance of School Desegregation, 1954-1972", unpublished MA Thesis, Department of Sociology, the University of Chicago.

Sheatsley, P.B.
1966 "White Attitudes toward the Negro", Daedalus, Vol. 95, No. 1, pp. 217-238.

Table 1: Attitude Towards Busing, by Race

	BUSING					
RACE	Approve	Disapprove	Strongly Disapprove	Total	N	Missing Data
White	6.6%	37.2%	56.2%	100.0%	697	30
Black	50.0	38.8	11.2	100.0	374	26
					1071	56

Chi-square 374.70 df 4 p 0.000

Goodman-Kruskal Tau: Busing .1378

Table 2: Attitude Towards White and Black Students Attending the Same Schools, by Race

	SCHOOLS					
RACE	Same	Unsure	Separate	Total	N	Missing Data
White	86.0%	4.7%	9.3%	100.0%	688	39
Black	93.4	3.7	2.9	100.0	390/1078	10/49

Chi-square 19.617 df 8 p .0119

Goodman-Kruskal Tau Schools .0105

Table 3: Attitude Towards Children Attending Schools Where a Few of the Children are of Opposite Race, by Race

	FEW OPPOSITE RACE				
RACE	No Objection	Object	Total	N	Missing Data
White	91.6%	8.4%	100.0%	722	5
Black	95.0	5.0	100.0	397/1019	3/8

Chi-square 5.010 df 2 p .0817

Goodman-Kruskal Tau Few .0044

Table 4: Attitude Towards Children Attending School Where Half of the Children are of Opposite Race, by Race

	HALF OPPOSITE RACE				
RACE	No Objection	Object	Total	N	Missing Data
White	64.5%	35.5%	100.0%	715	12
Black	91.7	8.3	100.0	396/1111	4/16

Chi-square 98.332 df 2 p 0.000

Goodman-Kruskal Tau Half .0880

Table 5: Attitude Towards Children Attending School Where More Than Half of the Children are of Opposite Race, by Race

	MORE THAN HALF OPPOSITE RACE				
RACE	No Objection	Object	Total	N	Missing Data
White	33.0%	67.0%	100.0%	648	79
Black	84.0	16.0	100.0	375/1023	25/104

Chi-square 250.05 df 2 p 0.000

Goodman-Kruskal Tau More than .2430

Table 6: Observed Frequencies in the 5-Way Table, Busing by Race by Children by Age by Education

				BUSING		
RACE	Children	AGE	EDUCATION	Approve	Disapprove	Strongly Disapprove
White	Yes	<40 yrs	≤12 yrs	8	29	66
			>12	2	16	17
		≥40	≤12	4	21	48
			>12	3	11	25
	No	<40	≤12	3	19	31
			>12	4	35	28
		≥40	≤12	5	51	79
			>12	5	11	33
Black	Yes	<40	≤12	33	31	11
			>12	10	11	3
		≥40	≤12	22	16	4
			>12	3	0[1]	2
	No	<40	≤12	21	8	3
			>12	11	10	4
		≥40	≤12	37	35	5
			>12	12	8	2
				183	312	301

[1] 0.5 was added to all cells before carrying out the log-linear analysis.

Table 7: Models Fit to the Data of Table 6 (Busing by Education by Age by Children by Race) and Assessments of Their Fit

Model	Marginals Fit*	Likelihood Ratio X^2	df	Signif.
(1)	B,EARC	319.59	30	0.000
(2)	EB,EARC	318.22	28	0.000
(3)	AB,EARC	317.03	28	0.000
(4)	RB,EARC	34.48	28	0.1855
(5)	CB,EARC	317.61	28	0.000
(6)	EB,AB,RB,CB,EARC	28.31	22	0.1657
(7)	EAB,RB,CB,EARC	17.15	20	>.500
(8)	ERB,AB,CB,EARC	24.71	20	0.2179
(9)	ECB,AB,RB,EARC	27.88	20	0.1125
(10)	ARB,EB,CB,EARC	25.46	20	0.1845
(11)	ACB,EB,RB,EARC	27.14	20	0.1315
(12)	RCB,EB,AB,EARC	26.56	20	0.1484
(13)	EAB,ERB,ECB,ARB,ACB,RCB,EARC	7.99	10	>.500
(14)	EARB,ECB,ACB,RCB,EARC	7.36	8	0.4986
(15)	EACB,ERB,ARB,RCB,EARC	6.91	8	>.500
(16)	ERCB,EAB,ARB,ACB,EARC	6.39	8	>.500
(17)	ARCB,EAB,ERB,ECB,EARC	5.15	8	>.500
(18)	EAB,RB,EARC	19.44	22	>.500
(19)	EAB,CB,EARC	301.13	22	0.000
(20)	EAB,EARC	302.82	24	0.000
(21)	EAB,ERB,EARC	15.58	20	>.500

*In model descriptions the variables are: B - Busing, E - Education, A - Age, R - Race, C - Children in public schools. The symbol "XY" means the model is constrained to reproduce the observed relation between variables "X" and "Y".

Table 8: Chi-square on the Difference Between the Fit of Selected Models in Table 7

Models	Terms Tested	Difference in Likelihood Ratio X^2	Difference in df	Signif.
(2) - (1)	EB	1.37	2	>.250
(3) - (1)	AB	2.56	2	>.250
(4) - (1)	RB	285.11	2	<.001
(5) - (1)	CB	1.98	2	>.250
(7) - (6)	EAB	11.16	2	<.005
(8) - (6)	ERB	3.60	2	>.100
(9) - (6)	ECB	0.43	2	>.750
(10) - (6)	ARB	2.85	2	>.100
(11) - (6)	ACB	1.17	2	>.500
(12) - (6)	RCB	1.75	2	>.250
(14) - (13)	EARB	0.63	2	>.500
(15) - (13)	EACB	1.08	2	>.500
(16) - (13)	ERCB	1.60	2	>.250
(17) - (13)	ARCB	2.84	2	>.100
(13) - (7)	--	9.16	10	>.500
(7) - (18)	--	2.29	2	>.250
(21) - (18)	--	3.86	2	>.100

Table 9: Observed Frequencies in the 5-Way Table, MIXING by Race by Children by Age by Education

RACE	CHILDREN	AGE	EDUCATION	MIXING: No objection	MIXING: Object
White	Yes	<40 yrs	≤12 yrs	25	73
			>12	13	21
		≥40	≤12	18	51
			>12	17	17
	No	<40	≤12	18	33
			>12	20	41
		≥40	≤12	44	83
			>12	21	5
Black	Yes	<40	≥12	59	16
			>12	21	3
		≥40	≤12	35	6
			>12	4	0[1]
	No	<40	≤12	21	9
			>12	17	25
		≥40	≤12	76	4
			>12	23	0
				432	387

[1] 0.5 was added to all cells before carrying out the log-linear analysis.

Table 10: Models Fit to the Data of Table 11 (MIXING by Education by Age by Children by Race) and Assessments of their Fit

Model	Marginals Fit*	Likelihood Ratio X^2	df	Signif.
(1)	M, EARC	251.90	15	0.000
(2)	EM, EARC	251.31	14	0.000
(3)	AM, EARC	239.96	14	0.000
(4)	RM, EARC	88.91	14	0.000
(5)	CM, EARC	250.69	14	0.000
(6)	EM, AM, RM, CM, EARC	64.97	11	0.000
(7)	EAM, RM, CM, EARC	50.31	10	0.000
(8)	ERM, AM, CM, EARC	46.67	10	0.000
(9)	ECM, AM, RM, EARC	62.32	10	0.000
(10)	ARM, EM, CM, EARC	51.62	10	0.000
(11)	ACM, EM, RM, EARC	53.89	10	0.000
(12)	RCM, EM, AM, EARC	59.82	10	0.000
(13)	EAM, ERM, ECM, ARM, ACM, RCM, EARC	12.44	5	0.0289
(14)	EARM, ECM, ACM, RCM, EARC	12.30	4	0.0151
(15)	EACM, ERM, ARM, RCM, EARC	7.07	4	0.1319
(16)	ERCM, EAM, ARM, ACM, EARC	7.29	4	0.1211
(17)	ARCM, EAM, ERM, ECM, EARC	7.71	4	0.1027
(18)	EARM, EACM, ERCM, ARCM, EARC	0.05	1	>.500
(19)	EACM, ERCM, ARCM, EARC	0.07	2	>.500
(20)	EACM, ARCM, EARC	3.35	4	0.4998
(21)	EACM, ARM, RCM, EARC	10.85	5	0.0542
(22)	ARCM, EAM, ECM, EARC	10.16	5	0.0705

*In model descriptions the variables are: M - MIXING; E - education; A - age; R - race; C - children in public schools. The symbol "XY" means the model is constrained to reproduce the observed relation between variables "X" and "Y".

Table 11: Chi-square Values on the Difference Between the Fit of Selected Models in Table 12

Models	Terms Tested	Difference in Likelihood Ratio X^2	Difference in df	Signif.
(2) - (1)	EM	0.60	1	>.250
(3) - (1)	AM	11.95	1	<.001
(4) - (1)	RM	163.00	1	<.001
(5) - (1)	CM	1.22	1	>.250
(7) - (6)	EAM	14.66	1	<.001
(8) - (6)	ERM	18.30	1	<.001
(9) - (6)	ECM	2.65	1	>.100
(10) - (6)	ARM	13.35	1	<.001
(11) - (6)	ACM	11.08	1	<.001
(12) - (6)	RCM	5.15	1	>.010
(14) - (13)	EARM	0.14	1	>.500
(15) - (13)	EACM	5.37	1	>.010
(16) - (13)	ERCM	5.15	1	>.010
(17) - (13)	ARCM	4.73	1	>.025
(18) - (19)	--	0.02	1	>.750
(19) - (20)	--	3.28	2	>.100

WHITE FLIGHT AND CENTRAL CITY LOSS: APPLICATION OF AN ANALYTIC MIGRATION FRAMEWORK[1]

William H. Frey, Center for Demography and Ecology
Department of Sociology, University of Wisconsin-Madison

This paper[2] utilizes an analytic migration framework to assess the aggregate impact of selected community-level factors on white population losses experienced in central cities of large metropolitan areas. The framework parameterizes analytically distinct components of local and long distance migration streams which contribute directly to central city population change. Each component can be specified as a function of community-level attributes which are relevant to the explanation of specific in- and out-migration streams.

In this application, previously advanced racial and nonracial attributes of central cities and their surrounding suburbs are used to estimate framework components based on 1970 census data for white movement streams associated with the central cities of large SMSAs. These estimates are then used to ascertain the impact that the central city racial composition exerts on net white out-migration from selected cities. The data demonstrate that the aggregate impact of racially linked "white flight" has been minimal.

I. Analytic Migration Framework

The framework was developed in order to analyze population change in both the city and suburbs of a metropolitan area through community determinants of movement streams that contribute directly to such change (see Frey, 1977a). Because each contributing stream responds to different sets of community attributes, the framework can be used to assess the net-migration consequences of city, suburb, and metropolitan attributes which influence movement levels in one or more streams. The core of the framework consists of a series of stream-specific parameters which can be linked to a demographic accounting equation, Through this linkage, relationships can be specified between community attributes, stream movement levels and aggregate population change in cities and suburbs.

The Framework Parameters

Each of the framework parameters are associated with one of the following movement streams:

I. Intrametropolitan City-to-Suburb or Suburb-to-City Mobility Streams
II. In-migration Streams to Cities or suburbs from outside the SMSA
III. Out-migration Streams from Cities or Suburbs to places outside the SMSA

The framework assumes that city and suburban population change are linked to population change at the metropolitan level and that the streams listed above represent all avenues whereby the city or suburb population is affected by movement within and from outside the metropolitan area. With one exception, the framework parameters associated with each stream represent rates which are applied to various "at risk" populations of residents and movers. These are listed in Figure A.

Beginning with the intrametropolitan city-to-suburb stream (stream IA), the rate at which a city resident will move to the suburbs during an interval is defined as the product of the parameters i_c and $p_{c\rightarrow s}$. This separation of parameters is prompted by empirical studies which show that residential mobility results from two major stages of decision-making -- the decision to move (made by a resident) and the choice of destination (made by the mover), and that each stage is influenced by different causal factors (Butler et al., 1969; Speare, Goldstein and Frey, 1975). Therefore, the i_c parameter denotes the rate at which a city resident will move anywhere within the SMSA, and the $p_{c\rightarrow s}$ parameter denotes the rate at which a city-origin mover will relocate in the suburbs. As will be demonstrated below, this distinction permits the analyst to causally relate different sets of community attributes to each stage of the mobility process. In a similar manner, the rate at which a suburban resident will move to the city (stream IB) is defined as the product of framework parameters i_s and $p_{s\rightarrow c}$.

In-migration to the central city or suburbs from outside the SMSA (streams IIA and IIB) is also seen to be the product of two framework parameters. For each stream, the number of in-migrants rather than the rate of in-migration is specified. In-migrants to the central city are defined as the product of parameters M_o and $p_{o\rightarrow c}$. M_o denotes the number of in-migrants to the SMSA as a whole, and $p_{o\rightarrow c}$ denotes the rate at which SMSA in-migrants locate in the central city. This separation of parameters is justified on the basis of findings that long-distance migrants are initially attracted to metropolitan-wide economic or labor market attributes (Lansing and Mueller, 1967). The city or suburb residential location within the metropolitan area then becomes a secondary decision for SMSA in-migrants which is made on the basis of different sets of factors.

Finally, only one framework parameter is associated with out-migration streams from metropolitan cities and suburbs (streams IIIA and IIIB).

The Demographic Accounting Equation

The framework parameters are linked to a demographic accounting equation which allows their effects to be translated into aggregate changes in city and suburb population sizes during an interval. If one begins with P_c^t, the city population at time t, and P_s^t, the suburb population at time t, it is possible to compute the city and suburb populations of age n and over at time t+n using the relationships in Figure B.

By employing these relationships, the migration framework can be used to relate community attributes to aggregate population change in central cities and suburbs. The key mechanisms for the analysis are the framework parameters which are assumed to be causally related to various attributes. More specifically, each

Figure A: Movement Streams and Associated Framework Parameters

IA - INTRAMETROPOLITAN CITY-TO-SUBURB MOBILITY

i_c MOBILITY INCIDENCE RATE OF CITY RESIDENTS
The rate at which city residents* move anywhere within the SMSA between t, t+n

$p_{c\to s}$ SUBURB DESTINATION PROPENSITY RATE OF CITY MOVERS
The rate at which city-origin movers relocate to a suburb destination between t, t+n

IB - INTRAMETROPOLITAN SUBURB-TO-CITY MOBILITY

i_s MOBILITY INCIDENCE RATE OF SUBURB RESIDENTS
The rate at which suburb residents* move anywhere within the SMSA between t, t+n

$p_{s\to c}$ CITY DESTINATION PROPENSITY RATE OF SUBURB MOVERS
The rate at which suburb-origin movers relocate to a city destination between t, t+n

IIA - IN-MIGRATION TO THE CITY FROM OUTSIDE THE SMSA

M_o MIGRATION INTO THE SMSA
Total number of migrants into the SMSA between t, t+n

$p_{o\to c}$ CITY DESTINATION PROPENSITY RATE OF IN-MIGRANTS
The rate at which SMSA In-Migrants relocate to a city destination between t, t+n

IIB - IN-MIGRATION TO THE SUBURBS FROM OUTSIDE THE SMSA

M_o MIGRATION INTO THE SMSA
Total number of migrants into the SMSA between t, t+n

$p_{o\to s}$ SUBURB DESTINATION PROPENSITY RATE OF IN-MIGRANTS
The rate at which SMSA In-Migrants relocate to a suburb destination between t, t+n

IIIA - OUT-MIGRATION FROM THE CITY TO OUTSIDE THE SMSA

$m_{c\to o}$ OUT-MIGRATION INCIDENCE RATE OF CITY RESIDENTS
The rate at which city residents migrate out of the SMSA between t, t+n

IIIB - OUT-MIGRATION FROM THE SUBURBS TO OUTSIDE THE SMSA

$m_{s\to o}$ OUT-MIGRATION INCIDENCE RATE OF SUBURB RESIDENTS
The rate at which suburb residents migrate out of the SMSA between t, t+n

*residents who do not out-migrate between t, t+n

Figure B: Demographic Accounting Equations

$$(1)\quad P_{c*}^{t+n} = sP_c^t - sP_c^t m_{c\to o} - s(P_c^t - P_c^t m_{c\to o})i_c p_{c\to s} + s(P_s^t - P_s^t m_{s\to o})i_s p_{s\to c} + sM_o p_{o\to c}$$

$$(2)\quad P_{s*}^{t+n} = sP_s^t - sP_s^t m_{s\to o} - s(P_s^t - P_s^t m_{s\to o})i_s p_{s\to c} + s(P_c^t - P_c^t m_{c\to o})i_c p_{c\to s} + sM_o p_{o\to s}$$

where:

P_{c*}^{t+n} = city population age n and over at time t+n

P_{s*}^{t+n} = suburb population age n and over at time t+n

s = survival rate for movers, migrants, or nonmovers

P_c^t = city population at time t

P_s^t = suburb population at time t

framework parameter can be expressed as a function of a number of community attributes which serve as independent variables. For example:

$$i_c = f(X_j)$$

where X_j denotes one of k community attributes which are related to the residential mobility incidence rate of city residents.

The other framework parameters can be specified as functions of the same or different attributes. After the parameters have been specified as functions of relevant community attributes, the demographic accounting equations can be used to assess the aggregate impact of an attribute (or combination of attributes) on population change in an individual city or suburb during an interval t, t+n.

II. Application to Central City "White Flight"

In this application of the analytic framework, we are interested in ascertaining the extent to which the size of the city's Black population influences aggregate white loss due to the selective suburban relocation of residential (intrametropolitan) movers, and the suburban destination choices of in-migrants to the metropolitan area.

The motivation for this investigation draws from an earlier study we had undertaken to assess the relative importance of both racial and nonracial influences on recent white city-to-suburb movement in large SMSAs (Frey, 1977b). Based on a cross-sectional analysis of movement streams in 39 SMSAs during the 1965-70 period, our findings indicated that racial influences did not predominate. Significant racial desegregation in central city schools and the occurrence of racial disturbances during the period contributed little to the explanation of city-to-suburb white flight, while ecological features of the SMSA and city-suburb fiscal disparities proved to be important determinants. One racial factor -- the percent of the central city population which was Black -- did influence white out-movement, particularly in non-Southern cities, and prevented us from dismissing racial factors completely as flight determinants.

The present analysis represents a somewhat restricted application of the framework in the sense that community attributes will only be assessed as determinants of the destination propensity parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$. This focus on the destination propensity parameters only can be justified on the basis of our earlier finding that the racial factor, percent city Black, influences white city-to-suburb movement primarily through the city-suburb destination choices of city-origin movers, and only minimally through the mobility incidence of city residents (denoted by framework parameter i_c) (Frey, 1977b). It is also consistent with studies of residential mobility motivations which indicate that the decision to move is affected less by "white flight" considerations than by the family's need to make housing adjustments coincident with changes in its size and composition (Rossi, 1955; Speare, Goldstein and Frey, 1975).

One further restriction will be the focus only on movement-induced changes to the size of the white city population, thus disregarding the effects of fertility and mortality on aggregate change.

The Data

The data for the investigation are taken from the Census subject report Mobility in Metropolitan Areas (U.S. Bureau of the Census, 1973) which classifies 1970 residents of cities and suburbs of the 65 largest SMSAs according to their 1965 residence locations, and from which it is possible to compute white (nonBlack) population and framework parameters for the 1965-70 interval that are necessary to pursue this analysis. These data will be used for two purposes: (a) to specify framework parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ as functions of community attributes; and (b) to calculate the increment to white city population loss in selected SMSAs that can be attributed to the community attribute, percent city Black. Specification of the destination propensity rates as functions of community attributes will be accomplished in cross-sectional multiple regression analyses, using as cases, the 39 SMSAs which were examined in the earlier study.

In order to calculate incremental white population change in selected SMSAs that is associated with different values of $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ using equation (1) in Figure B, it is necessary to obtain actual values for the remaining framework and population parameters in that equation. These can also be computed from the 1970 Census subject report, although for this purpose it is useful to rearrange the terms of that equation (see footnote to Table 1).

Specifying Framework Parameters

The community attributes that are used to estimate destination propensity parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ constitute those racial and nonracial attributes which proved to be the most important determinants of white city-to-suburb movement in our earlier study. These attributes and their abbreviations are as follows.[3]

- BLK -- Percent City Black
- CIT -- City Share of SMSA Population
- EDX -- Suburb/City Educational Expenditures Per Capita (x 100)
- TAX -- Suburb/City Tax Revenues Per Capita (x 100)
- CRM -- City Crime Rate
- PSD -- Postwar Suburban Development
- CMT -- City-Suburb Commuters
- CTA -- Central City Age: The number of years between the census year when the city first attained a population of 50,000 and the year 1970
- SRG -- Southern Region: (South=1, Other Regions=0)
- S×B -- Interaction of SRG and BLK

We now proceed to specify the framework parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ as functions of the community attributes just presented in regression analyses. Each parameter is regressed on all of the attributes for the 39 SMSAs that form the basis of this investigation. The resulting equations appear as follows:

$$p_{c\to s} = +.3164 +.0024\,BLK -.0076\,CIT +.0008\,EDX -.0012\,TAX +.0003\,CRM +.0038\,PSD +.0024\,CMT +.0006\,CTA +.0411\,SRG -.0006\,S\times B \qquad (3)$$

$$R^2 = .92$$

$$p_{s\to c} = +.0671 -.0004\ BLK +.0059\ CIT +.0003\ EDX -.0007\ TAX -.0008\ CRM -.0013\ PSD +.0027\ CMT -.0012\ CTA -.0492\ SRG +.0019\ S\times B \quad (4)$$

$R^2 = .84$

$$p_{o\to c} = +.0249 -.0038\ BLK +.0113\ CIT +.0004\ EDX -.0012\ TAX +.0001\ CRM -.0018\ PSD +.0036\ CMT -.0007\ CTA -.0606\ SRG +.0029\ S\times B \quad (5)$$

$R^2 = .93$

It is difficult to evaluate the relative importance of each attribute from the unstandardized coefficients presented here. It is, nevertheless, apparent that the percent city Black increases the suburb propensity of city movers and decreases the city propensity of suburb movers and SMSA in-migrants. Each of these effects is greatly moderated in Southern cities.

The Aggregate Impact on White City Loss

We move on to the major aim of this analysis: to ascertain the aggregate impact on white city loss which can be attributed to the city's Black population size as it affects the destination choices of white residential movers and SMSA in-migrants. This aggregate impact will be assessed in three SMSAs: Cleveland, Dayton, and Dallas. Each of these had a fairly sizeable percentage of Blacks in the central city at the beginning of the migration interval: 33% for Cleveland, 26% for Dayton, and 22% for Dallas.

Presented in Table 1 are the 1965-70 population and framework parameters for Cleveland, Dayton, and Dallas which are necessary to estimate P_{c*}^{1970} for each city. The values for parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ are estimated from equations (3), (4), and (5) based on actual values for the community attributes shown in Table 2. The values for the remaining framework and population parameters were computed from actual mobility and population data for the SMSAs reported in the 1970 census.

To assess the aggregate impact of BLK, the following strategy will be taken: First, we assume various actual and hypothetical numbers of Blacks in each city for 1965. Second, we translate these actual and assumed numbers into values of Percent City Black (BLK). Third, we compute parameters $p_{c\to s}$, $p_{s\to c}$, $p_{o\to c}$ from the actual and hypothetical values of BLK using equations (3), (4), and (5). Fourth, we compute 1970 white city population figures (P_{c*}^{1970}) based on actual and hypothetical values of $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ using the demographic accounting equation (1). The latter figures will allow us to compare the aggregate changes to each city's white population which would have resulted from different racial mixes in the city at the beginning of the movement interval.

The results of this analysis appear in Table 3. For each of the three SMSAs, the following series of assumptions is made about the number of central city Blacks in 1965: (A) the actual number of Blacks, (B) a 50 percent increase in the actual number, (C) a 25 percent increase in the actual number, (D) a 25 percent decrease in the actual number, and (E) a 50% decrease in the actual number. Shown in column (1) are the corresponding values of BLK which are used to estimate the destination propensity parameters in columns (2) through (4). The final three columns display results of the computations using the demographic accounting equation (1): the white city population age 5 and over (column 5), the difference from the actual total (column 6), and the percent difference from the actual total (column 7).

As our review of equations (3), (4), and (5) suggested, an increase in the Percent City Black is associated with a net decrease in the white population. Yet the level of impact resulting from the drastic differences in the number of city Blacks is not substantial in any of the three cities. This effect is extremely small in Dallas -- resulting in part from the lesser influence of Percent City Black in Southern SMSAs. Clearly, the aggregate "flight" impact of the central city racial composition -- as transmitted through the destination choices of local movers and in-migrants -- is slight, over a five-year migration interval.

III. Use of the Framework in "White Flight" Research

The investigation undertaken here represents an initial step toward a causal analysis of white central city population change utilizing the analytic migration framework. This framework, which we have described in more detail elsewhere (Frey, 1977a), allows the researcher to identify city, suburb, and metropolitan determinants of movement streams which contribute directly to population change in the central city. Using this framework in conjunction with readily available census data, it is possible to calculate incremental changes in a city's population associated with specific community attributes that serve as determinants of one or more movement streams. In this manner, the framework can be employed to establish causal relationships between community attributes, stream movement levels, and aggregate population change in the central city, over the course of a migration interval.

In the present application, we focused our attention on one causal attribute -- city racial composition -- as it affects white central city change through the selective destination choices of white intrametropolitan movers, and white in-migrants to the metropolitan area. Based on aggregate movement data from selected large SMSAs, our findings indicate that such effects were minimal over the 1965-70 interval. Hence, not only does the city's racial composition play a relatively minor role in explaining white movement from the city to the suburbs (Frey, 1977b), but the total impact of its influence on aggregate white city loss seems also to be exceedingly small, at least in the short-run.

Although restricted in its focus to one causal attribute and three framework parameters, this application of the analytic framework serves to illustrate its utility in an investigation of central city "white flight" determinants. In future reports, we plan to extend our causal analysis of white population loss beyond this re-

strictive focus in order to incorporate a greater number of community attributes as causal factors, and to provide a more refined assessment of "flight" consequences for central city change.

FOOTNOTES

[1] This research is supported by grant No. 1 R01 HD-1-666-01, "Migration and Redistribution: SMSA Determinants," from the Center for Population Research of the National Institute of Child Health and Human Development.

[2] A more extented treatment appears in Center for Demography and Ecology Working Paper 77-27 University of Wisconsin-Madison.

[3] Fuller definitions and rationale for these factors appear in Frey (1977b).

REFERENCES

Brown, Lawrence A., and Eric G. Moore. 1970. The Intra-Urban Migration Process: A Perspective. Geografiska Annaler 52B: 1-13.

Butler, Edgar W., et al. 1969. Moving Behavior and Residential Choice -- A National Survey. National Cooperative Highway Research Program Report No. 81. Washington, D.C.: Highway Research Board, National Academy of Sciences.

Frey, William H. 1977a. "Population Movement and City-Suburb Redistribution: An Analytic Framework." Presented at the 1977 Meetings of the Population Association of America, St. Louis, Missouri. (Working Paper 77-15, Center for Demography and Ecology, University of Wisconsin-Madison.)

_____. 1977b. "Central City White Flight: Racial and NonRacial Causes." For presentation at the 1977 Meetings of the American Sociological Association, Chicago, Illinois. (Discussion Paper No. 420-77, Institute for Research on Poverty, University of Wisconsin-Madison.)

Lansing, John B., and Eva Mueller. 1967. The Geographic Mobility of Labor. Ann Arbor: Institute for Social Research.

Rossi, Peter H. 1955. Why Families Move. New York: The Free Press.

Shryock, Henry S. Jr., and Jacob S. Siegel. 1973. The Methods and Materials of Demography. Washington, D.C.: U.S. Bureau of the Census.

Simmons, James W. 1968. Changing Residence in the City: A Review of Intraurban Mobility. Geographic Review 58: 622-651.

Speare, Alden Jr., Sidney Goldstein, and William H. Frey. 1975. Residential Mobility, Migration and Metropolitan Change. Cambridge, Massachusetts: Ballinger Publishing Company.

U.S. Bureau of the Census. 1973. Census of Population: 1970. Subject Reports Final Report PC(2)-2C. Mobility for Metropolitan Areas. Washington, D.C.: U.S. Government Printing Office.

Table 1: Population and Framework Parameters for the 1965-70 interval[a] used as inputs to Equation (1)[b]

SMSAs	$s(P_c^{1965} - P_c^{1965}m_{c\to o})$	$s(P_c^{1965} - P_c^{1965}m_{c\to o})i$	$p_{c\to s}$	$s(P_s^{1965} - P_s^{1965}m_{s\to o})i_c$	$p_{s\to c}$	sM_o	$p_{o\to c}$
Cleveland	435015	195720	.422	261724	.101	141307	.228
Dayton	167571	89756	.507	120206	.080	101326	.189
Dallas	445161	204591	.342	158816	.214	261200	.453

[a] Framework parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ are estimated from equations (3), (4), and (5) in the text based on actual community attributes (see Table 2). The other population and framework parameters are computed from the 1970 Census subject report Mobility in Metropolitan Areas (U.S. Bureau of the Census, 1973).

[b] Equation (1) can be rewritten as:

$$P_{c*}^{t+n} = s(P_c^t - P_c^t m_{c\to o}) - s(P_c^t - P_c^t m_{c\to o})i_c p_{c\to s} + s(P_s^t - P_s^t m_{s\to o})i_s p_{s\to c} + sM_o p_{o\to c}$$

where t = 1965, n = 5, and s represents the appropriate survival rate for each mover, migrant, or nonmover group.

Table 2: Community Attributes used to Estimate Framework Parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ for 1965-70 Interval in Cleveland, Dayton, and Dallas SMSAs

Community Attributes[a]	Cleveland	Dayton	Dallas
BLK	33.1	26.0	22.3
CIT	41.0	32.1	57.0
EDX	92.9	103.6	109.9
TAX	77.7	54.2	50.7
CRM	59.3	66.1	59.7
PSD	58.8	62.4	71.3
CMT	23.9	21.7	10.9
CTA	100.0	80.0	60.0
SRG	0.0	0.0	1.0
SxB	0.0	0.0	22.3

Table 3: The Effects of Actual and Hypothetical Numbers of City Blacks in 1965 on Migration Framework Parameters $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ during the 1965-70 Interval, and on the 1970 City White Population Age 5 and over, in Cleveland, Dayton, and Dallas SMSAs

Assumed Number of City Blacks in 1965:	BLK Value[a] (1)	1965-70 Parameter Values[b] $p_{c\to s}$ (2)	$p_{s\to c}$ (3)	$p_{o\to c}$ (4)	1970 City White Population Age 5 and Over: Population Size[c] (5)	Difference from (A) (6)	Pct Difference from (A) (7)
Cleveland SMSA							
A. Actual Number	33.1	.422	.101	.228	411153	--	--
B. Increase by 100%	49.7	.462	.095	.165	392848	-18305	- 4.5
C. Increase by 50%	42.6	.445	.098	.192	400701	-10452	- 2.5
D. Decrease by 50%	19.8	.391	.106	.279	425751	+14598	+ 3.6
E. No Blacks	0.0	.344	.114	.354	447570	+36417	+ 8.9
Dayton SMSA							
A. Actual Number	26.0	.507	.080	.189	150777	--	--
B. Increase by 100%	41.2	.544	.074	.131	140959	- 9818	- 6.5
C. Increase by 50%	34.5	.528	.076	.157	145304	- 5473	- 3.6
D. Decrease by 50%	14.9	.481	.084	.231	157884	+ 7107	+ 4.7
E. No Blacks	0.0	.446	.090	.288	167482	+16705	+11.1
Dallas SMSA							
A. Actual Number	22.3	.342	.214	.453	527378	--	--
B. Increase by 100%	36.4	.367	.235	.440	522362	- 5016	- 1.0
C. Increase by 50%	30.1	.356	.225	.446	524619	- 2759	- 0.5
D. Decrease by 50%	12.5	.324	.199	.461	530828	+ 3450	+ 0.7
E. No Blacks	0.0	.302	.180	.472	535268	+ 7890	+ 1.5

[a] BLK is computed for each assumed number of city Blacks in 1965 as: $\frac{\text{(assumed number of 1965 city Blacks)}}{\text{(assumed number of 1965 city Blacks + actual number of 1965 city whites)}} \times 100$

[b] Computed from equations (3), (4), and (5) based on column (1) value of BLK and the actual values of CIT, EDX, TAX, CRM, PSD, CMT, CTA, SRG, and SxB which appear in Table 2.

[c] Computed from equation (1) [see footnote to Table 1], based on values of $p_{c\to s}$, $p_{s\to c}$, and $p_{o\to c}$ in columns (2), (3), and (4) and on actual values for the other framework parameters which appear in Table 1.

BLACK-WHITE HOUSING QUALITY DIFFERENTIALS IN THE UNITED STATES, 1970[1]

M. E. El Attar, S. El Attar, W. Frese and M. S. Al-Marayati
Mississippi State University

INTRODUCTION

It is well known that adequate housing in the United States is not available for every household. For the United States as a whole, 63.1% of the housing was rated as "sound"[2] in 1950, this increased to 74.0% by 1960 (U. S. Bureau of the Census, 1953: 5 and 1962: 5). The 1970 Census of Housing did not provide data on housing quality comparable to that reported in 1950 and 1960. Specifically, the 1970 quality of housing was based on adequacy of plumbing facilities and crowdedness (instead of plumbing and structural condition). According to the 1970 Census of Housing, the category that was almost comparable to 1950 and 1960 provided an adequate housing percentage of 86.2 (U.S. Bureau of the Census, 1973: 7). Regional breakdown of the data reveals wide differences among regions and divisions of the United States with regard to housing quality. Table 1 provides the contrasting data. The states having the largest percentage of "sound" quality housing were in the Pacific Division, with the states in the East South Central Division having the smallest percent of sound quality housing (Table 1).

OBJECTIVE OF THE STUDY

The objective of this study is to determine if there was a significant difference in the quality of housing for whites and blacks of equal income in 1970. Many changes occured in the area of housing quality and racial discrimination after the 1960 Census. New housing policies and programs were implemented and laws against discrimination were more strictly enforced. However, in spite of these developments there still seem to exist differences in quality of housing for blacks and whites. It is the purpose of this paper, then, to test the hypothesis that in 1970 there were differences in the quality of housing units rented or owned by blacks and whites in the same income brackets.

DATA AND ANALYSIS

The data for this study were obtained from the Census of Housing: 1970 Metropolitan Housing Characteristics, United States and Regions. Specifically, the following tables were used in the compilation of data.

1. "Income in 1969 of families and primary individuals in owner and renter occupied housing units: 1970," for whites and Negroes.
2. "Plumbing facilities by persons per room for owner and renter occupied housing units: 1970," for whites and Negroes.

Definitions and explanations of basic concepts used in this study are those adopted by the U. S. Bureau of the Census as stated in the 1970 Census of Housing reports.[3]

Descriptive Statistics

The percent of good quality housing units and poor quality housing units occupied by white and black owners and renters at each level of income are presented in Table 2.[4] The figures in Table 2 are based upon the total number of occupied housing units in the 1970 Census for owners and renters, total and blacks. The Census provides statistics for 63,445,192 housing units, of these, 39,885,545 were occupied by owners and 23,559,647 by renters. The total number of housing units occupied by blacks was 4,646,701, of these, 2,567,761 were occupied by owners and 2,078,940 by renters.

In general, Table 2 shows a direct relationship between income level and quality of housing for owners and renters for both blacks and whites. It should be pointed out, however, that the percentage of whites tends to be greater than blacks in the case of owners occupying "good" quality housing while the reverse is true for renters of this same quality. Conversely, in the case of owners and renters occupying housing of "poor" quality, the percentages for blacks tend to be higher than whites. This implies that blacks occupy housing units of less quality than those of whites.

Inferential Statistical Analysis

In order to ascertain the significance of the observed differences between blacks and whites, a two-way analysis of variance was applied to the data in Table 2. Table 3 summarizes these results. The null hypothesis tested stated that there were no differences in the 1970 quality of housing units rented or owned by blacks and whites in the same income category. The hypothesis was rejected in three cases (good quality, owner; good quality, renter; and poor quality, owner) and accepted in the case of renters occupying poor quality housing units.

To assess the contribution of income to these significant housing differences, an analysis of variance for contrasts was performed. Table 4 provides a summary of the significant results which indicate differences in housing quality for blacks and whites in the following income levels: (1) over $6,999 for plumbing facilities in owned housing units; (2) above $6,999 for ratio of crowdedness in owned housing units; and (3) below $7,000 for plumbing facilities in rented housing units.

TABLE 1. PERCENTAGE OF HOUSING OF SOUND QUALITY IN REGIONS AND DIVISIONS OF THE UNITED STATES, 1950, 1960 AND 1970

Region and Division	1950[a]	1960[b]	1970[c]
Northeast	78.0	80.6	95.6
New England	73.9	78.7	94.9
Middle Atlantic	79.2	81.2	95.8
North Central	61.2	74.5	92.9
East North Central	66.1	77.1	93.8
West North Central	50.9	68.6	90.7
South	44.6	63.4	73.3
South Atlantic	48.8	67.2	73.3
East South Central	32.7	53.0	61.2
West South Central	47.6	64.8	81.2
West	78.4	81.8	94.2
Mountain	62.3	74.8	87.6
Pacific	83.6	83.9	96.7
U. S. Total	63.1	74.0	86.2

Source: U. S. Bureau of the Census, 1953, 1962, and 1973, Tables 1.

[a]"With hot running water, private toilet and bath, and not dilapidated."
[b]"Sound with all plumbing facilities."
[c]"With all plumbing facilities."

TABLE 2. PERCENTAGE OF WHITE AND BLACK OWNER AND RENTER OCCUPIED HOUSING UNITS OF GOOD QUALITY AND POOR QUALITY BY INCOME LEVELS, UNITED STATES: 1970

Income $1,000	Good				Poor			
	Owner		Renter		Owner		Renter	
	White	Black	White	Black	White	Black	White	Black
<2	5.03	5.01	4.41	10.68	0.77	2.39	0.77	4.31
2-	2.79	2.34	2.25	4.98	0.27	0.80	0.28	1.54
3-	2.67	2.32	2.28	4.83	0.21	0.66	2.23	1.29
4-	2.61	2.31	2.30	4.39	0.17	0.53	0.18	0.92
5-	2.88	2.44	2.53	4.27	0.16	0.42	0.16	0.67
6-	3.15	2.52	2.58	3.83	0.13	0.33	0.13	0.47
7-	11.69	7.14	6.90	8.04	0.28	0.54	0.23	0.72
10-	17.18	7.23	6.15	5.32	0.18	0.24	0.11	0.27
15-	11.33	3.72	2.65	1.58	0.07	0.05	0.03	0.06
25+	3.56	0.56	0.63	0.22	0.02	0.03	0.01	0.02

Source: Compiled and computed from U. S. Bureau of the Census, 1972, Tables A-4 and A-14.

TABLE 3. ANOVA FOR QUALITY OF HOUSING UNITS OCCUPIED BY BLACKS AND WHITES, OWNERS AND RENTERS: 1970

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Squares	Computed F
		Good Quality, Owner		
Due to income	230.370	9	25.597	4.074**
Due to color	37.264	1	37.264	5.931*
Residual	56.547	9	6.283	
Total	324.181	19		
		Good Quality, Renter		
Due to income	91.405	9	10.156	4.392**
Due to color	11.951	1	11.951	5.168*
Residual	20.813	9	2.313	
Total	124.169	19		
		Poor Quality, Owner		
Due to income	3.492	9	0.388	3.449*
Due to color	0.696	1	0.696	6.182*
Residual	1.013	9	0.113	
Total	5.201	19		
		Poor Quality, Renter		
Due to income	12.044	9	1.339	1.939
Due to color	1.885	1	1.885	2.731
Residual	6.211	9	0.690	
Total	20.140	19		

*Significant at 0.05 **Significant at 0.025

TABLE 4. ANOVA FOR THE QUALITY OF HOUSING UNITS OCCUPIED BY BLACKS AND WHITES OF DIFFERENT INCOME LEVELS, RENTERS AND OWNERS, 1970

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Squares	Computed F*
	Less than $2,000			
Plumbing, renters	6,017,209.00	1,16	6,017,209.00	5.79
	$2,000 - $2,999			
Plumbing, renters	973,182.25	1,16	973,182.25	6.01
	$3,000 - $3,999			
Plumbing, renters	814,506.25	1,16	814,506.25	8.07
	$4,000 - $4,999			
Plumbing, renters	500,910.06	1,16	500,910.06	8.06
	$5,000 - $5,999			
Plumbing, renters	314,440.56	1,16	314,440.56	7.81
	$6,000 - $6,999			
Plumbing, renters	160,200.06	1,16	160,200.06	6.25
	$7,000 - $9,999			
Plumbing, owners	1,148,648.01	1,16	1,148,648.01	8.11
Room density, owners	6,109,181.70	4,16	1,527,295.42	10.78
	$10,000 - $14,999			
Plumbing, owners	611,325.62	1,16	611,325.62	17.61
Room density, owners	2,458,895.80	4,16	614,723.95	17.70
	$15,000 - $24,999			
Plumbing, owners	3,622,360.60	1,16	3,622,360.60	58.04
Room density, owners	14,132,362.00	4,16	3,533,090.50	56.61
	$25,000 or more			
Plumbing, owners	562,500.00	1,16	562,500.00	45.52
Room density, owners	2,186,341.50	4,16	546,585.38	44.24

*All computed F ratio values are at least significant at the .05 level. For a complete analysis of variance see Al-Marayati, 1977, Tables III through XII, pp. 28-37.

FOOTNOTES

[1]The research on which this paper is based is a part of MAFES Population Project No. 4004. The authors gratefully acknowledge the contribution of Dr. Rose M. Rubin.

[2]In 1950 and 1960, quality of housing was measured by its "structural condition" ("'dilapated'" or "'not dilapated'") and "plumbing facilties." The term "sound" was introduced to express this quality. See (U. S. Bureau of the Census, 1954:XIV and XV and 1962); and (Bird, 1973:3). According to these criteria, the term "sound housing" refers to "housing with no defects or only slight defects" and "which has hot and cold running water, flush toilet and bathtub (or shower) inside the structure for the exclusive use of the occupants" (U. S. Bureau of the Census, 1962:XXIV).

[3]A Housing Unit is a house, an apartment, a group of rooms, or a single room occupied or intended for occupancy as separate living quarters. Separate living quarters are those in which the occupants do not live and eat with any other persons in the structure and which have either (1) direct access from the outside of the building or through a common hall or (2) complete kitchen facilities for the exclusive use of the occupants." Income "is the sum of the amounts reported" in 1969 "for wages and salary income, net self-employment income, Social Security or railroad retirement income, public assistance or welfare income, and all other income." Persons per room is the number of persons in a housing unit divided by the number of rooms in the unit. Plumbing Facilities denotes "units which have hot and cold piped water inside the structure as well as a flush toilet and a bathtub or shower inside the structure for the exclusive use of the occupants of the unit." "Race refers to the race of the head of the household occupying the housing unit." Housing quality in this study, was classified as good if the units had all "plumbing facilities and a ratio of 1.0 or more rooms per person. Housing units were classified as poor quality housing units if they lacked one or more enumerated plumbing facilities and/or had a ratio of less than 1.0 rooms per person."

[4]The percentages in the table are obtained by dividing the figures for whites and blacks by their relevant total.

REFERENCES

Al-Marayati, Mahdi
1977 "Housing Quality Differentials Between Whites and Blacks in the United States: 1970." Unpublished M.A. Thesis, Mississippi State University.

Bird, Ronald
1973 Inadequate Housing and Poverty Status of Households, Areas Served by Farmers Home Administration Programs, 1970, By States. Statistical Bulletin No. 520, June. Washington, D.C.: Rural Development Service, U. S. Department of Agriculture.

U. S. Bureau of the Census
1953 County and City Data Book, 1952, (A Statistical Abstract Supplement). Washington, D.C.: U. S. Government Printing Office.

U. S. Bureau of the Census
1962 County and City Data Book, 1962, (A Statistical Abstract Supplement). Washington, D.C.: U. S. Government Printing Office.

U. S. Bureau of the Census
1972 Metropolitan Housing Characteristics. Final Report HC(2). Washington, D.C.: U. S. Government Printing Office.

U. S. Bureau of the Census
1973 County and Data Book, 1972, (A Statistical Abstract Supplement). Washington, D.C.: U. S. Government Printing Office.

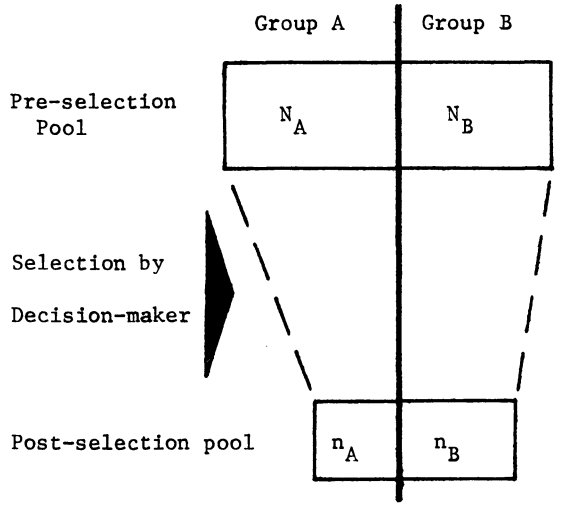
STATISTICAL MODELING TO SUPPORT A CLAIM OF INTENTIONAL DISCRIMINATION

James W. L. Cole, Pittsburgh, PA
David C. Baldus, University of Iowa

In the context of discrimination cases the courts have acknowledged that "Figures speak, and when they do, courts listen."[1] Indeed, conclusions drawn from statistical analyses often form an important and valued element of the evidence supporting claims of discrimination against minorities. But when one reads the cases in which statistics have been applied, one observes that the courts have left unresolved a number of methodological issues. Our purpose here is to examine some issues that arise in a particular class of discrimination cases and to pursue their resolution. The cases to be considered have the following defining properties in common:

1. The plaintiffs belong to a group (say, Group A) which is constitutionally or statutorily protected.
2. The practice under challenge is a selection process which assigns to each candidate for selection one of two outcomes, namely selection or rejection. Thus candidates are effectively winnowed out in a process represented schematically as in Figure 1 which highlights
   a) the pre-selection pool of applicants containing Group A (protected minority) and Group B (majority or other) components.
   b) the similarly constituted post-selection pool of those selected, and
   c) the decision maker, the individual or group actually making the choices.

FIGURE 1



3. The essence of the challenge is that the decision-maker was covertly applying a policy that used group membership to influence the applicants' selection chances and hence put Group A applicants at a disadvantage relative to others. However, those elements of the policy governing the selection process that are open to public view do not explicitly refer to group membership or any equivalent criterion as a factor influencing chances of selection.
4. The policy governing the selection process has left the decision maker ample room to use personal discretion in making choices, in that choices are not determined substantially by qualification criteria that are open to public view.

Selection processes susceptible to challenges with these properties will include those found in the substantive areas of employment selection, promotion, school admissions, some instances of criminal sentencing, and jury selection. The very early landmark case Yick Wo v. Hopkins[2] involved such a claim, and a long and rich line of challenges against jury selection processes provide numerous other examples.[3]

What is being claimed in cases having these properties is intentional discrimination of the type prohibited in particular under the due process and equal protection clauses of the 14-th Amendment of the U. S. Constitution. In two recent opinions, namely Washington v. Davis[4] and Arlington Heights v. Metropolitan Housing Corporation[5], the Supreme Court has made it clear that, to be successful, such challenges raised on constitutional grounds require proof of two facts. These are the existence of discriminatory impact, relative disadvantage falling to the plaintiff as a result of the suspect practice, and the existence of an intent to discriminate underlying the practice.

The dual nature of this requirement has been unclear in many decisions against jury selection systems because in these decisions a single piece of evidence has been found to establish a prima facie case of intentional discrimination. Specifically, the plaintiffs have prevailed on showing that over a suitably long period of time citizens with the same group membership have been substantially or consistently under-represented on the jury lists. However, in its discussion of requirements for proof of intentional discrimination in Castaneda v. Partida, the Supreme Court observed that the sufficiency of a single piece of evidence in such cases simply reflects the fact that the one fact can have two implications: the substantial under-representation speaks directly to the question of discriminatory impact while at the same time suggesting by its magnitude the presence of illicit motive.[6]

Selection processes can be classified usefully according as the decision-maker's choices are or are not guided by consideration of overt and verifiable qualification criteria. If the decision-

maker's choices are influenced in part by such qualifications we will call the process a 'guided' discretionary process. In such processes, because equal protection doctrine is concerned with the equal treatment of equally situated candidates, such qualification variables that legitimately divide the candidates into equivalence classes of 'equal situation' must be taken into account before either a relevant measure of impact can be obtained or a valid inference of motive can be drawn.

If no such qualifications are cited, we will call the process 'purely' discretionary. With respect to those aspects of the governing policy that are open to public view, and hence for purposes of measuring the discriminatory impact of the discretionary aspects of the policy, each candidate who exceeds a basic threshold level of eligibility can be considered to be as qualified for selection as any other such candidate.

Because purely discretionary processes will be easier to model, we will consider them first. Fortunately, they do provide a suitable context for discussing interesting questions.

Modeling in Purely Discretionary Processes

For challenges of purely discretionary selection processes the four frequencies, N_A, N_B, n_A, and n_B found in the accompanying four-fold table contain the information usually considered necessary to establish a prima facie case of either discriminatory impact or discriminatory intent.

Table 1

Group	Numbers Selected	Numbers Rejected	Pre-selection Pool Totals
Group A	n_A	$N_A - n_A$	N_A
Group B	n_B	$N_B - n_B$	N_B
Totals	n	N - n	N

Usually the information in these quantities is summarized by a number or statement that compares one measure that reflects how the minority group was actually treated with a corresponding measure constructed to show how the group should have or would have been treated absent any discriminatory behavior. The following measures of actual treatment are frequently used:

a) the selection rate (or pass rate), $P_A = n_A/N_A$,

b) the rejection rate (or fail rate), $1 - P_A$,

c) the inverse selection rate, $1/P_A$,

d) the minority representation rate in the post-selection pool, $r_A = n_A/n$

(to be compared with $R_A = N_A/N$, the representation rate in the pre-selection pool), or even

e) the actual number of minority candidates chosen, n_A.

These rates or numbers can be compared in a variety of ways. Usually comparisons are made in terms of arithmetic differences or ratios, but other formulas have been suggested.[7] The courts have had some problems with the variety of measures available, the most obvious being a lack of consistency and direction in the choice of formulas for summarizing the numerical information bearing on the issue at hand. However, this lack of direction is not in itself the most troublesome aspect. As the courts move away from cases with clear-cut factual bases and encounter those that are closer, they are more inclined to compare the numbers in the case at hand with those of precedent cases.[8] To do this without adopting some function of the tetrad (n_A, n_B, N_A, N_B) that evaluates the evidence in terms of a single number is to trust the reliability of subjective judgment. On the other hand, any function chosen to evaluate the tetrad in the context of a particular question of fact should have an essentially monotonic (strictly) increasing relationship with the actual legal significance of the tetrad as it relates to the question of fact. For example, if in a given situation an arithmetic difference of 10 percentage points between R_A and r_A is much more strongly suggestive of unlawful motive when $R_A = 11\%$ than when $R_A = 91\%$, the blind dependence on $R_A - r_A$ to support an inference of motive will invite the drawing of erroneous conclusions.

We propose that a primary determinant in choosing a measure should be the purpose of the measure--whether it is intended to measure discriminatory impact or to support an inference of motive. In measuring discriminatory impact, we are inquiring as to the relative harm done to members of the minority group as they go through the selection process, and this suggests that the underlying modeling be motivated by a utility theoretic approach. On the other hand, in inferring motive, we are seeking to identify an aspect of the decision-maker's behavior and principles of behavioral modeling should dominate.

Measuring the Discriminatory Impact. In a purely discretionary selection system, each Group A applicant entering the process can be said to have the same probability of being selected, say P_A, as any other such applicant. If we assume further that each candidate has the same (positive or negative) utility, say u, of being selected and utility 0 (zero) of being rejected[9], then the expected utility for a minority candidate is clearly

$$E_A(U) = u \cdot P_A .$$

Finally, assuming that in the absence of discrimination minority candidates would have the same probability of selection as that which the majority candidates have, say P_B, the expected utility for a minority candidate would be

$$E^*(U) = u \cdot P_B .$$

Therefore a measurement of harm should clearly be based on some comparison between $u \cdot P_A$ and $u \cdot P_B$. However, to be consistent with the principles of utility theory, a given arithmetic shortfall in the expected utility must represent the same degree of harm to the applicant regardless of the value of the expected utility that would obtain in a non-discriminatory process. That is to say, the appropriate measure of harm is the arithmetic difference

$$E^*(U) - E_A(U) = u(P_B - P_A).$$

Finally, since the minority and majority probabilities of selection are estimated without statistical bias by the corresponding observed selection rates, these results clearly suggest that a measure of discriminatory impact based on the difference between selection rates is preferable.[10]

A significant exception to this argument applies when the selection process under challenge is that of selecting the venire from which a jury will be chosen when that jury is to decide the fate of a minority criminal defendant. For in this situation, the courts must be concerned primarily with the impact on the rights of the defendant, not those of the prospective veniremen. Moreover, in this situation, the composition of the post-selection pool (that is, the venire from which the jury will ultimately be chosen) is the only aspect of the venire-selection process that bears on the defendant's rights. Consequently, while the difficulty of modeling the subsequent jury selection may make the application of the utility-theoretic argument impractical and its results of doubtful acceptability to the courts, it is clear that the minority representation rates that result from the venire selection process are the pivotal quantities in establishing the impact on the defendant's rights.

Modeling to Infer Discriminatory Motive. When we turn to the goal of modeling to infer motive, the focus of the model shifts from the treatment of the applicant to the behavior of the decision-maker. The first step in this modeling process is to pin-point as nearly as possible the kind of non-discriminatory selection mechanism the decision-maker might have used or claims to have used. The second is to adopt a model that appears to simulate that mechanism adequately. The third is to consider how discriminatory behavior might manifest itself in the context of the given selection mechanism. The fourth is to decide how that form of discriminatory behavior is to be incorporated into the model. Finally, we apply the conclusions drawn in the first four steps to the pursuit of our current objective, namely to choose a measure best suited to reflecting behavior suggestive of a discriminatory motive in the eyes of the court.

For example, the selection of prospective jurors from the citizens in a district is often supposedly done by application of an explicitly random mechanism. The appropriate model for such a process will be more or less obvious depending on the complexity of the mechanism and the care taken to adhere to it. For example, in straightforward situations the model of simple random sampling from the pool of all citizens meeting specified eligibility requirements may suffice, but when allowances are made for various forms of hardship, the model may require modification.

In the context of such a process, two forms of discrimination would seem to be most likely. The first is the ever-present possibility that the number of Group A applicants was held below some tacit quota. The second is the exclusion from consideration of some fixed portion, consisting say of πN_A individuals, from the pool of eligible Group A candidates.

If the process is being influenced by a desire to limit Group A selections to a quota, it is more likely that the decision-maker is concerned with controlling the size of the Group A representation rate than the Group A selection rate. Hence this desire is more likely to be reflected in the representation rate than the selection rate. Specifically, if the nominal model for the process is one of simple random sampling, the distribution of the representation rate would be expected to differ from the nominal binomial or hypergeometric distribution in two ways. Naturally, the mean of the distribution will be reduced from the expected N_A/N. But also the distribution would likely be truncated, especially at the upper end. This suggests that when results from several applications of the suspect process are available, as is often the case in jury selection challenges, the entire empirical distribution of representation rates may be relevant.[11]

If the Group A participation in the selection process is limited by exclusion of a fixed proportion of eligible candidates, an estimate of the excluded proportion based on the representation rates has been suggested.[12] This estimate is found by simple algebra to be

$$\hat{\pi} = \frac{R_A - r_A}{R_A(1 - r_A)}$$

Purely random mechanisms are less common in employment selection, where typically some form of evaluation or ranking will be used. If all the evaluation criteria are strictly subjective and hence inaccessible for verification or challenge, the process will still fall in the 'purely discretionary' class. However, the five-step process of analysis and modeling should reflect the dependence on the criteria if possible.

Such selection mechanisms provide more interesting modeling challenges. The particular model chosen will depend, first, on the nature of the criteria on which the informal evaluation is based; second, on what can be postulated as reasonable distributions for these criteria in the populations of candidates; and finally, on the kind of reasoning that was used to combine these criteria into a single score.

In order to develop an illustration, albeit more valuable for insights produced than for realism, suppose the following. First, for each candidate there exists a vector of qualification scores $\underset{\sim}{X}_{ij}$ $(i = 1, \cdots, N_j;\ j=A,B)$. Second, these vectors are p-variate normally distributed within groups--

$$\underset{\sim}{X}_{ij} \sim N_p(\mu_j, \Sigma)$$

--with the same covariance matrix but possibly different means. And finally, the decision-maker looks at these criterion scores, constructs some weighted sum of them, say

$$Y_{ij} = \underset{\sim}{\lambda}_j' \underset{\sim}{X}_{ij},$$

and selects those candidates for whom Y_{ij} exceeds a cut score y_j^*.[13]

This model will reflect an absence of intentional discrimination only if $\underset{\sim}{\lambda}_A = \underset{\sim}{\lambda}_B$ and $y_A^* = y_B^*$. Any difference between the weight vectors or the cut scores would indicate that group membership was being used to influence the decisions.

According to this model the weighted sum Y_{ij} will be normally distributed,

$$Y_{ij} \sim N(\underset{\sim}{\lambda}_j' \mu_j;\ \underset{\sim}{\lambda}_j' \Sigma \underset{\sim}{\lambda}_j \equiv \sigma^2(\underset{\sim}{\lambda}_j)).$$

Thus, the probability that a candidate i, being randomly drawn from group j, will meet the standard for selection is

$$P(Y_{ij} \geq y_j^*) = \Phi\left[\frac{\underset{\sim}{\lambda}_j' \mu_j - y_j^*}{\sigma(\underset{\sim}{\lambda}_j)}\right].$$

Therefore we find that $\Phi^{-1}P(Y_{ij} \geq y_j^*)$ will be in the form of a linear model--

$$\Phi^{-1} P(Y_{ij} \geq y_j^*) = \alpha_0 + \alpha_1 Z_{ij}$$

--where $Z_{ij} = 1$ for candidates in Group A, and 0 otherwise; and the unknown coefficients have the following form:

$$\alpha_0 \equiv \frac{\underset{\sim}{\lambda}_B' \underset{\sim}{\mu}_B}{\sigma(\underset{\sim}{\lambda}_B)} - \frac{y_j^*}{\sigma(\underset{\sim}{\lambda}_B)}$$

$$\alpha_1 \equiv \frac{1}{\sigma(\underset{\sim}{\lambda}_B)} \underset{\sim}{\lambda}_B' (\underset{\sim}{\mu}_B - \underset{\sim}{\mu}_A) + \left(\frac{1}{\sigma(\underset{\sim}{\lambda}_A)} \underset{\sim}{\lambda}_A' - \frac{1}{\sigma(\underset{\sim}{\lambda}_B)} \underset{\sim}{\lambda}_B'\right) \underset{\sim}{\mu}_A + \frac{y_A^*}{\sigma(\underset{\sim}{\lambda}_A)} - \frac{y_B^*}{\sigma(\underset{\sim}{\lambda}_B)}$$

In this linear model, the parameter α_1 contains all the evidence of discriminatory behavior. This suggests using an appropriate estimate of α_1 as our measure from which to infer motive. From the methodology of probit analysis[14] we find the maximum likelihood estimate of α_1 to be $\Phi^{-1}(p_B) - \Phi^{-1}(p_A)$, where p_A and p_B are the observed selection rates for the two groups.

In practice, this measure would likely be resisted as unfamiliar and based on unverifiable assumptions. Thus we seek a more familiar and intuitive measure that would produce similar results in the case-to-case comparisons. We turn first to the logit function, $L(p) \equiv \ln\left(\frac{p}{1-p}\right)$. Since, as can be shown,

$$0.83 < \frac{(L(P_B)-L(P_A))/1.9}{\Phi^{-1}(P_B) - \Phi^{-1}(P_A)} < 1.16 \quad (.05 \leq P_A, P_B \leq .95)$$

the difference $L(p_B) - L(p_A)$ suggests itself as an alternative with the virtue of being easily explained in terms of betting odds. Finally, we note that if p_A and p_B are both small, the simple ratio p_B/p_A will give case-to-case comparisons most consistent with those of $\Phi^{-1}(p_B)-\Phi^{-1}(p_A)$ among the measures now commonly used.

As a last word it should be noted that α_1 will reflect the arguably innocent effect of the group mean differences in qualification score inextricably confounded with the influences of intentional discrimination. This fact would suggest an obvious defense to the claim, and hence it raises legal and methodological questions that are serious and most interesting, but beyond the scope of this discussion.

Modeling in Guided Discretionary Processes

In the following discussion of guided discretionary processes we will assume that there is no challenge to the legitimacy of the overt qualification criteria guiding the decision-maker, but that in making the final selection decisions, the selector has covertly used group membership as a factor. Thus it is alleged that, within the 'equivalence classes' of candidates defined by virtue of being similarly situated with respect to the overt criteria, the selector operated to the disadvantage of the Group A members. Consequently, we shift our focus, both in measuring discriminatory impact and in detecting evidence of unlawful intent, to comparisons of treatment observed within the equivalence classes.

Measuring the Discriminatory Impact. Let $\underset{\sim}{W}_{ij}$ denote the vector of overt qualifications and let $P_j(\underset{\sim}{w})$ denote the probability of selection for a group j candidate whose $\underset{\sim}{W}_{ij}$ vector equals $\underset{\sim}{w}$. Then invoking the same assumptions of utility structure and the same argument as in the previous section, the suggested measure of impact for a minority candidate with qualification vector equal to $\underset{\sim}{w}$ would be an appropriate estimate of the difference $P_B(\underset{\sim}{w}) - P_A(\underset{\sim}{w})$.

This difference can be estimated in two general ways: (1) directly from selection outcomes observed for groups of candidates having $\underset{\sim}{W}_{ij} = \underset{\sim}{w}$, if there exist such groups of sufficient size, or alternatively, (2) adopting models for $P_A(\underset{\sim}{w})$ and $P_B(\underset{\sim}{w})$ as functions of $\underset{\sim}{w}$ and estimating the parameters therein by appropriate methods.

If the latter approach is chosen, it will again be apparent that the model to use will again depend on the particular situation. But in this application of modeling, the final result will be judged purely on the reliability of the estimates that it provides for $P_B(\underset{\sim}{w}) - P_A(\underset{\sim}{w})$ as indicated in part by measures of goodness-of-fit of the model. If such measures indicate the need for a model in which $P_B(\underset{\sim}{w}) - P_A(\underset{\sim}{w})$ varies with $\underset{\sim}{w}$, the lack of a single number summarizing the magnitude of the impact for the whole case will clearly complicate case-to-case comparisons. These complications are intrinsic to situations in which some portions of Group A may have been treated more adversely than others, and it would be unwise to confine the choice of models artificially to those which have an additive term corresponding to group membership.

Modeling to Infer Discriminatory Motive. When modeling to lay the basis for an inference of motive in a guided discretionary selection process, the same multi-step analysis of the mechanism should be applied as for a purely discretionary process, with the recognition that observed values for the overt qualification criteria will permit empirical goodness of fit testing of some aspects of the resulting model.

To illustrate, we will extend the informal evaluation model of the completely discretionary process section to incorporate the qualification variables $\underset{\sim}{W}_{ij}$. Thus we will assume that adjoining the q variable vector $\underset{\sim}{W}_{ij}$ to $\underset{\sim}{X}_{ij}$ produces a vector distributed (p+q)-variate normally--

$$\begin{pmatrix} \underset{\sim}{W}_{ij} \\ \underset{\sim}{X}_{ij} \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \underset{\sim}{\nu}_j \\ \underset{\sim}{\mu}_j \end{pmatrix}; \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right), \begin{pmatrix} i = 1,\cdots,N_j \\ j = A,B \end{pmatrix}$$

And, as before, we assume that the decision-maker constructs some weighted sum of all the qualification scores, say

$$Y_{ij} = \underset{\sim}{\delta}_j' \underset{\sim}{W}_{ij} + \underset{\sim}{\lambda}_j' \underset{\sim}{X}_{ij}$$

and selects those candidates for whom Y_{ij} exceeds y_j^*. In the context of this model, intentional discrimination would be implied by any of the following findings:

$$\underset{\sim}{\lambda}_A \neq \underset{\sim}{\lambda}_B, \quad \underset{\sim}{\delta}_A \neq \underset{\sim}{\delta}_B, \quad \text{or} \quad y_A^* \neq y_B^* .$$

According to this model, the distribution of Y_{ij} conditioned on a particular value for $\underset{\sim}{W}_{ij}$ is normal with mean $\underset{\sim}{\delta}_j' \underset{\sim}{w} + \underset{\sim}{\lambda}_j' [\underset{\sim}{\mu}_j + B(\underset{\sim}{w} - \underset{\sim}{\nu}_j)]$ and variance $\sigma^2(\lambda_j) \equiv \underset{\sim}{\lambda}_j' \Sigma_{22 \cdot 1} \underset{\sim}{\lambda}_j$, where $B \equiv \Sigma_{21}\Sigma_{11}^{-1}$ and $\Sigma_{22 \cdot 1} \equiv \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Thus the conditional probability that a candidate i in group j will be selected given $\underset{\sim}{W}_{ij} = \underset{\sim}{w}$ is

$$P(Y_{ij} \geq y_j^* | W_{ij}=w) = \Phi\left(\frac{w'(\delta_j + B'\lambda_j) + \mu_j^{*\prime}\lambda_j - y_j^*}{\sigma(\lambda_j)}\right),$$

where $\mu_j^* = \mu - B\nu_j$ is the vector of X intercepts in the regression functions on w. Again we find that $\Phi^{-1}P(Y_{ij} \geq y_j^* | W_{ij} = w)$ will be in the form of a linear model--

$$\Phi^{-1}P(Y_{ij} \geq y_j^* | W_{ij}=w) = \alpha_0 + \alpha_1 Z_{ij} + \alpha_2' w + \alpha_3' Z_{ij} w,$$

--where Z_{ij} is again an indicator function for membership in Group A, and the unknown coefficients have the following form:

$$\alpha_0 = \frac{\lambda_B'\mu_B^* - y_B^*}{\sigma(\lambda_B)}$$

$$\alpha_1 = \frac{-1}{\sigma(\lambda_B)}\lambda_B'(\mu_B^* - \mu_A^*) + \left(\frac{1}{\sigma(\lambda_A)}\lambda_A' - \frac{1}{\sigma(\lambda_B)}\lambda_B'\right)\mu_A^* - \left(\frac{y_A^*}{\sigma(\lambda_A)} - \frac{y_B^*}{\sigma(\lambda_B)}\right)$$

$$\alpha_2 = \frac{1}{\sigma(\lambda_B)}(\delta_B + B'\lambda_B)$$

$$\alpha_3 = \left(\frac{1}{\sigma(\lambda_A)}\delta_A - \frac{1}{\sigma(\lambda_B)}\delta_B\right) + B'\left(\frac{1}{\sigma(\lambda_A)}\lambda_A - \frac{1}{\sigma(\lambda_B)}\lambda_B\right)$$

In this model, the parameters α_1 and α_3 contain the evidence of discriminatory behavior, and are the center of our interest. Obtaining estimates of α_1 and α_3 using the methods of probit analysis will present real problems unless the values of W for the candidates are concentrated in not too small groups at just a few different points. However, even for the case where the values of W are spread thinly over many points a method described by Walker and Duncan[16] offers a method of estimating the parameters, provided that we are willing again to substitute the logit function for the inverse normal probability integral function.

Non-zero values in α_3 clearly indicate intentional discrimination in the form of differential recognition of qualifications in different groups. However, α_1 is again a sum reflecting both innocent and suspect influences and raising the same legal and methodological questions alluded to earlier.

In closing, we note that the complexity of having to consider legitimate qualification variables again results in substantial difficulty in defining a single summary measure to permit simple case-to-case comparison.

Footnotes

[1] Brooks v. Beto, Federal Reporter, 2d Series 366: 1 at p. 9 (1966).

[2] United States Reports 118: 356 (1886).

[3] For a recent example see Castaneda v. Partida, Supreme Court Reporter 97: 1272 (1977)

[4] Supreme Court Reporter 96: 2040 (1976)

[5] Supreme Court Reporter 97: 555 (1977)

[6] Supreme Court Reporter 97: 1272 at pages 1279-80. This distinction between the functions that evidence of under-representation on jury panels can fulfill had been drawn earlier in Swain v. Alabama (United States Reports 380: 202 at pp.208-9).

[7] One measure frequently suggested is the proportional shortfall in the representation rate, $(R_A - r_A)/R_A$. Another is the estimate of an excluded proportion discussed subsequently. See text at note 13.

[8] For example note comparisons made in Casteneda v. Partida (see note 3 above) at page 1281.

[9] This assumption is not critical for this analysis, for which each Group A candidate is assumed to have the same probability of selection. Eliminating this assumption would change the form but not the import of the argument.

[10] The persistence of u as a factor in $E^*(U)-E_A(U)$ raises a legitimate concern about the comparison of measures of impact for selection processes involving very different rewards. Thus the courts might well be more concerned with a seven percentage point difference between death sentencing rates applied to equally situated murder convicts than for a ten percentage point difference in hiring rates among equally qualified candidates. However, the appearance of u should not invalidate comparisons between results from situations in which the utility constants are similar.

[11] A discussion of hypothesis tests sensitive to under dispersed distributions of representation rates suggestive of quota limiting is found in M. O. Finkelstein, "The application of statistical decision theory to jury discrimination cases." Harvard Law Review 80: 338 (1966) at P. 365ff.

[12] J. Kirk in R. H. Amidon, et al., Mexican-Americans and Administration of Justice in the Southwest. (A report prepared for and issued through the U. S. Commission on Civil Rights)(1970) at pp. 132-3.

[13] The change in assumptions from fixing the number chosen to fixing the threshold of selection is made at some cost in credibility but with a large gain in simplicity of the argument.

[14] See, for example, D. J. Finney, Probit Analysis, Cambridge University Press (1964) at pp. 48-51.

[15] There are pitfalls in using p_B/p_A when both P_B and P_A are small as the value of the ratio will be very sensitive to sampling variability of p_A and to errors in the definition of the candidate pool for Group A. The first difficulty can be controlled by construction of an approximate confidence interval for P_B/P_A and the use of the most conservative value in the interval. The second requires close inspection of the assumptions and data gathering for the pre-selection pools.

[16]S. H. Walker and D. B. Duncan. "Estimation of the probability of an event as a function of several independent variables." Biometrika 54: 167 (1967).

DYNAMIC MODELING OF NEIGHBORHOOD TRANSITION

L. Sanathanan, W. O'Neill, and J. McDonald
University of Illinois at Chicago Circle

Neighborhood transition is a phenomenon that has been before the public for some time. This process can be described, locally at least, using the Chicago housing market. As pointed out by Berry (1976), between 1960 and 1970, 482,000 new housing units were built in the Chicago SMSA, while the number of households increased by only 285,000; a ratio of 1.7 new housing units for each new family. The effect of this was to promote a series of moves by upwardly mobile families to suburban areas which in turn exerted downward pressure on the prices of older housing units. The rapid occupancy changes in neighborhoods have had effects in certain areas in terms of the socioeconomic environment and school system. Measures leading to a quantitative understanding of the process of neighborhood change are obviously desirable. A policy decision regarding intervention into this process must be preceded by an understanding of the underlying mechanism and how certain policy decisions affect this mechanism.

This paper contains the results of a pilot study aimed at the development of a model for predicting the course and extent of neighborhood racial change. Such a predictive model is necessary for diagnosing whether a certain neighborhood needs intervention on the part of the housing authorities or the local government and for deciding on the extent of intervention. This will also enable us to classify neighborhoods according to their future prospects which should be an important consideration in the allocation of public community development funds.

In the present pilot study we focus on the Austin community in the city of Chicago. Austin is located on the west side of Chicago, east of the suburb of Oak Park. As a logical first step it was decided that an analysis of changes in real estate prices over time should form the basis for the study of neighborhood transition. For this analysis we turned to the data base that has been developed by Berry (1976), consisting of information for 30,000 transactions that took place in Chicago between 1968 and 1972. For each transaction, the location of the unit, its selling price, and the date of transaction are available. In addition, for many of the transactions involving single family dwellings, the assessed values of land and structure are provided. In order to detect possible price shifts, a regression analysis was performed under the assumption that the logarithm of deflated selling price is a polynomial function of time plus a linear function of the logarithms of assessed values of land and structure (to control for variations in housing characteristics and lot size). For the purpose of this regression, data pertaining to census tracts 2514 thru 2519 within the Austin area was used. Also, blocks were grouped into somewhat homogeneous clusters with each cluster being associated with a separate polynomial in time to account for the possibly different dynamic effects in the various clusters. Our analysis showed that prices did not undergo significant change when a neighbhood was experiencing racial transition. Numerous studies have failed to find significant price declines during and after racial transition. See Pascal (1970) for a discussion of such studies. Our efforts were then directed toward finding a significant measure of racial change in a neighborhood.

Our main finding is that the process of neighborhood transition is clearly reflected in the corresponding density of single family house transactions. This can be seen by examining some examples of frequency histograms as shown in Figure 1. Figure 1 shows the frequencies of transactions taking place in intervals of 200 days between 1968 January and 1971 December, within portions of tracts 2518, 2519, and 2520. We notice that each of these histograms starts at a low level, builds up to a peak, and recedes. There is, for instance, a flurry of selling activity in 2518C around the time point of 200 days, and almost no activity after 1000 days. We propose that what we are observing here is panic selling. Under normal circumstances, one would expect a constant turnover rate resulting in a uniform density of transactions. However in a panic market the percent per year sold to the emerging race in an area can be assumed to be proportional to the number of units presently held by the receding race. This is because panic selling requires a readily available housing stock in order to continue propagation. One would, of course, also expect normal selling activity, but the normal activity should be swamped by the panic effect. The panic selling process can be characterized mathematically using a difference equation whose solution yields a logistic type curve for the transaction volume. The mathematical details are given in the next section which also contains results of fitting such a logistic curve to observed data pertaining to a specific part of Austin. The fit is remarkably good as indicated by the R^2 value shown there.

The above theory is based on the simplifying assumption that the area under consideration is a homogeneous closed community in the sense that units within the community do not interact with those outside the community. In reality, there will be some edge effects, but as long as they are not pronounced, one can still detect the panic effect through a single well-defined peak. In most of the histograms that we examined this was clearly the case. In some instances we observed contaminated distributions resulting from imperfect groupings of blocks. The contamination was especially visible when we aggregated the data along the lines of census tracts. This suggests that strategic grouping of blocks is necessary.

We have presented a rationale and results that give us a basis for detecting and characterizing panic selling. This is only a preliminary step in our overall modeling effort. Our goal is to be able to predict for any given locality if and when it will experience a transition. The questions that need to be answered are: given the existing pattern, at what point in time (if ever) will the transition evidence itself in the locality in question and what would be the time course of the

transition? (The size of the peak measures both the speed of the transition and the total units susceptible to transition.) It is evident that the transition curve peaks at different times in the various localities we have considered, essentially giving rise to a travelling wave phenomenon. The next step would be to examine the different peaking times and magnitudes and to relate these to relevant geographic and socioeconomic aspects of the regions. The following concept borrowed from physics is helpful in this context. Consider two points in space and visualize a wave traveling from one point to the other. One characteristic of the wave is its velocity of propagation which is defined as distance/time taken for the wave to go from one point to the other. If we can estimate the velocity, then based on distance we would be able to predict the time between two peaks. Velocity would, of course, depend on a number of factors such as median education, median income, percent of foreign stock, percent of the population under 18, percent of units that are owner occupied, etc. (see Steinnes (1977)) for not only the immediate vicinity of the point under study, but also the intervening region (although not to the same extent as the former). It is clear, then, that velocity must be estimated as a function of these factors. This can be done using the data from the 1970 Census of Population and Housing, and transaction density curves of the type mentioned earlier. An estimate for the size of the peak can also be obtained in a similar manner. Finally, given a point in space for which a prediction has to be make, our strategy would be to consider the locality closest to it which has already undergone a transition and then to estimate the relevant parameters.

The Quantitative Model

We assume the neighborhood in question is composed of white households of number w and black households of number b. The total number of single unit households, N, is fixed so that

$$w + b = N.$$

Since little, if any, new single units were built in the period and locale of our data a constant population assumption is justified. If the neighborhood in question is sufficiently close to a region undergoing racial change we postulate the following model to hold

$$\frac{1}{b}\left(\frac{db}{dt}\right) = \beta w$$

where β is a rate constant. Using this equation and

$$w + b = N$$

$$\frac{db}{dt} \simeq b(t) - b(t-1)$$

yields the model

$$b(t) = (1+\beta N)b(t-1) - \beta b^2(t-1) \qquad (1)$$

We propose to estimate $1+\beta N$ and β from transaction histogram data. However the data used should be selected such that the region in question is fairly certain to have experienced a racial transition in the time period 1968-1972. Using 1970, SMSA data for the Austin area we plotted the percent black households versus census tract blocks. It was then evident that several block groupings would produce the conditions:

1) racial transition probably complete by 1968,
2) racial transition probably took place almost wholly in the period 1968-72,
3) racial transition probably would not occur in 1968-72.

Selecting an area satisfying condition 2 produced the transaction histogram of Figure 1 labeled "Blocks from 2518,19,20". To statistically test (1) we generated a sample time series, $b_s(t)$ as follows. Call the histogram values $h_s(t)$. Then our model requires

$$b_s(t) - b_s(t-1) = h_s(t).$$

Thus

$$b_s(1) = b_s(0) + h_s(0)$$

$$b_s(2) = b_s(0) + h_s(0) + h_s(1)$$

$$\vdots$$

$$b_s(t) = b_s(0) + \sum_{j=0}^{t-1} h_s(j).$$

Since the $h_s(j)$ are all available as the histogram the $b_s(t)$ sample series is known except for $b_s(0)$. We assumed $b_s(0)$ values about the same order as

$$\sum_{j=0}^{t-1} h_s(j),$$

that is, we took $b_s(0)$ = 50, 75, 100, 125, 150, 200 single family units. We then selected the $b_s(0)$ that yielded minimum standard deviation about the regression plane

$$b(t) = (1+\hat{\beta}\hat{N})b(t-1) - \hat{\beta}b^2(t-1)$$

where $\hat{\beta}$ and $\hat{N}$ are least squares estimates of β and N. With $R^2 = 0.984$ we get the equation

$$b(t) = \underset{(30.75)}{1.26}b(t-1) - \underset{(-5.25)}{0.00141}b^2(t-1) \qquad (2)$$

with the t values of the estimates in parentheses. Our estimates gave

$$\hat{\beta} = 0.00141$$

$$\hat{N} = 185.$$

A comparison of $b_s(t)$ for $b_s(0) = 100$ and a solution of (2) for $b(0) = 100$ is shown in Figure 2.

Note from Figure 2 the asymptotic character of the model response. The model predicts a total single family housing stock of 185 units which is the stable equilibrium value, say b_e, the model seeks. The equilibrium can be calculated as the solution of $b(t) \equiv b(t-1) = b_e$ in the equation

$$b_e = (1+\hat{\beta}\hat{N})b_e - \hat{\beta}b_e^2.$$

From 1970 SMSA data the total single family unit housing stock for the area used was 196 units.

Relationship to Tipping Theory

Since Grodzins (1957) first proposed the idea of a racial "tipping point," housing researchers have been trying to analyze racial change using this notion. Tipping can be defined as a distinct increase in the rate of racial transition which happens once the percentage black reaches a crucial level, the "tipping point". Grodzins proposed that the tipping point is 10% to 20%. The studies by Duncan and Duncan (1957) and Steines (1977) provide some confirmation for a tipping point. However, other researchers have failed to detect a tipping point (Rapkin and Grigsby (1960), Stinchcombe, et al (1969), and Wolf (1963)). For example, the studies by Wolf (1963) indicate a steady rate of racial change. Wolf's result is in contrast to our model which states

$$\frac{1}{w}\left(\frac{dw}{dt}\right) = -\beta b.$$

Our study tends to support the notion of tipping as defined above for a particular set of circumstances. As Rapkin and Grigsby (1960) have emphasized, the expectations which people have for a neighborhood are important determinants of behavior. In Chicago there is a long history of complete racial transition once the process begins. The area under investigation is located near areas that had undergone complete racial transition in the recent past. Thus, both blacks and whites probably formed similar expectations for the Austin area. This means that black demand in the area was strong because it was not expected to remain an all-white area. Given that blacks were willing to buy any house that was offered, the time path of racial transition follows the distribution of white tolerance levels for the percentage black. (See Schelling (1971) for more formal models of this type). This distribution of tolerance levels may or may not imply a tipping point greater than zero. However, our evidence of panic selling implies that the distribution of white tolerance levels was not uniform in the Austin area. Indeed unless one uses histograms that begin in time with values representative of normal transaction behavior it is questionable whether or not tipping phenomena will evidence itself.

We have tested our model to see if it could pick up tipping phenomena even though none of our histograms in Figure 1 clearly shows an "equilibrium" transaction rate preceeding the obvious panic selling rate. Specifically we tested the model

$$\dot{w}/w = -\beta(b - t_r)$$

where $t_r > 0$ is the tolerance level or threshold below which no panic selling occurs. Using $b + w = N$ and $\dot{w} = w(t) - w(t-1)$ gives the model

$$b(t) = [1+\beta(N+t_r)]b(t-1) - \beta b^2(t-1) - N\beta t_r. \quad (2)$$

Using the same data and $b_s(0)$ value as the testing of (1) we got ($R^2 = 0.984$) the equation

$$b(t) = \underset{(2.72)}{1.292}b(t-1) - \underset{(-0.89)}{.001522}b^2(t-1) - \underset{(-0.07)}{2.0915} \quad (3)$$

and the estimates

$$\hat{N} = 190, \ \hat{\beta} = 0.001522, \ \hat{t}_r = 7.23.$$

While the low t values on $\hat{\beta}$ and $N\beta t_r$ are disappointing, the sign and percent equivalent of

$$\hat{t}_r\left(= \frac{\hat{t}_r \times 100}{\hat{N}}\right) = 3.8\%$$

are what one would expect. Further, the autoregressive nature of b(t) would tend to lower these t values and compel one to judge the model fit more on the total R^2 and on how well the model predicts $b_s(t)$. On this score equation (3) essentially duplicates the data fit shown in Figure 2.

References

B. Berry, "Ghetto Expansion and Single-Family Housing Prices: Chicago, 1968-1972," Journal of Urban Economics, 3, (1976), 397-423.

O. Duncan and B. Duncan, The Negro Population of Chicago. Chicago: University of Chicago Press, 1957.

M. Grodkins, Metropolitan Segregation. Chicago: University of Chicago Press, 1957.

A. Pascal, "The Analysis of Residential Segregation," in J. Crecive, ed., Financing the Metropolis. Beverly Hills, California: Sage Publications, 1970.

C. Rapkin and W. Grigsby, The Demand for Housing in Racially Mixed Areas. Berkeley: University of California Press, 1960.

T. Schelling, "Dynamic Models of Segregation," Journal of Mathematical Sociology. 1, (1971), 143-186.

D. Steinnes, "Alternative Models of Neighborhood Change," Social Forces. 55, (1977), 1043-1058.

A. Stinchcombe, M. McDill, and D. Walker, "Is There a Racial Tipping Point in Changing Schools?", Journal of Social Issues. 25, (1969), 127-136.

E. Wolf, "The Tipping Point in Racially Changing Neighborhoods," Journal of the American Institute of Planners. 29, (1963), 217-222.

2518C

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
200.	26	**************************
400.	4	****
600.	3	***
800.	3	***
1000.	0	
1200.	1	*
1400.	0	
1600.	1	*
1800.	1	*

2519A

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
200.	15	***************
400.	24	************************
600.	16	****************
800.	16	****************
1000.	7	*******
1200.	4	****
1400.	7	*******
1600.	9	*********
1800.	1	*

2519C

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
0.	1	*
200.	25	*************************
400.	30	******************************
600.	4	****
800.	13	*************
1000.	11	***********
1200.	8	********
1400.	5	*****
1600.	4	****

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
0.	9	*********
200.	16	****************
400.	12	************
600.	6	******
800.	6	******
1000.	8	********
1200.	9	*********
1400.	5	*****
1600.	4	****
1800.	1	*

Blocks from 2518,19,20

FIGURE 1.

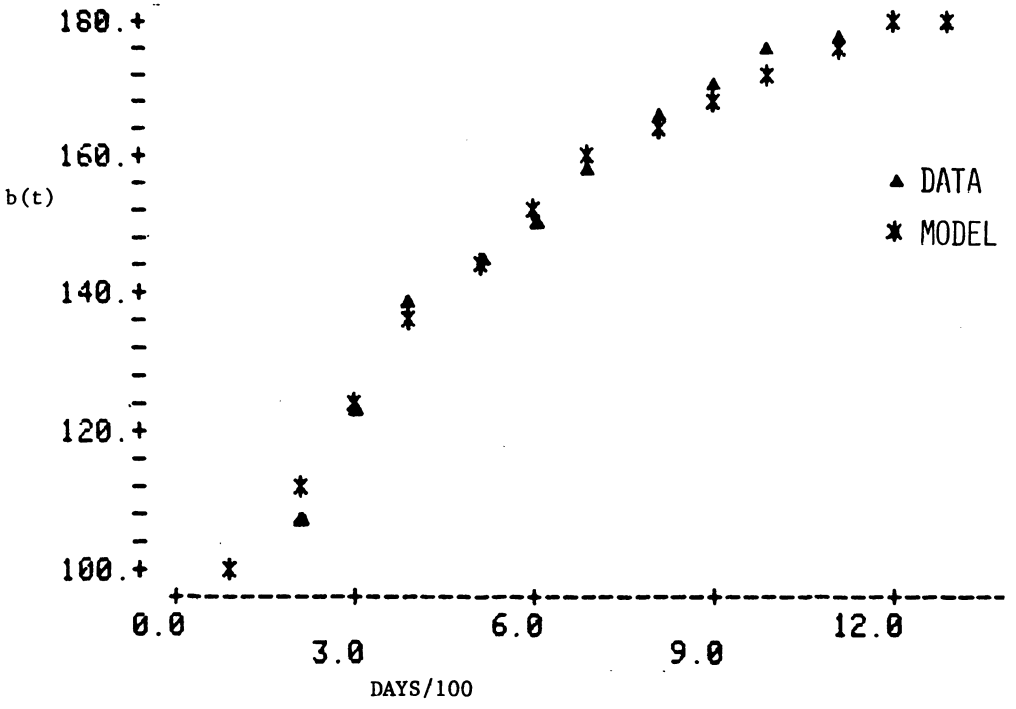


FIGURE 2.

IMPACT MODELS OF HIGHWAY CONSTRUCTION ON URBAN NEIGHBORHOODS: SOME EMPIRICAL RESULTS*

Ben-chieh Liu, Midwest Research Institute

I. Introduction

In recent years, there has been an upsurge of interest in problems of evaluating the impacts on urban neighborhoods of transportation development in general, and highway construction in particular. The issues related to transportation system impact have become more compelling than those related simply to the balance between the supply and demand for transportation services. In other words, people are becoming more concerned about the so-called "concomitant outputs" such as the tangible and intangible effects of the system on society and the environment (e.g., air pollution, noise, land utilization, urban sprawl, community life style, neighborhood cohesion, etc.) than about the "performance outputs" such as changes in travel times, volumes, costs and other objectives of the transportation system [12].

How may the relationships between the amount and distribution of travel and the social, economic, political and environmental impacts of transportation facilities and systems be identified, measured and evaluated? What specific changes can be recommended so that the performance outputs can be maximized and the adverse concomitant outputs minimized? What research is needed that would contribute to efficient and optimal decisions regarding the provision of transportation facilities and services in both the short and long run in urban and rural areas? Answers to these questions are of critical importance because any intelligent transportation decision requires the inputs from not only transportation engineers, architects and planners, but also from a variety of others such as ecologists, economists, sociologists, etc. In any decision regarding freeway construction, the questions are whether the benefits derived from the particular freeway are greater than the costs associated with the construction of the freeway--whether direct or indirect, tangible or intangible, social or private benefits and costs--and how they are measured.

The primary objectives of this paper are to empirically evaluate and to test the relevance and usefulness of some predictive models and to develop an alternative quality of life indicator model for neighborhood impact assessment. Empirical results on neighborhood life quality changes attributable to highway construction are also derived and discussed.

II. Impact Models of Highway Construction: An Evaluation

Three predictive methods--mobility index, social feasibility model, and neighborhood social interaction index--have been recently developed for predicting the highway construction effect on the neighborhood, each one has its weakness and strength and on the whole, none of them can adequately reflect the construction impacts on urban neighborhood life quality.

The mobility indicator developed by the California Division of Highways [3], in the form of a numerical index, was made up of the percentage of: (1) owner-occupied houses; (2) single family residences; and (3) people in the same house over 5 years. The California approach was extended and tested further by a Texas A&M study of 152 neighborhoods and 47 control neighborhoods in Austin, Dallas, and Houston [8,9]. Mobility Index (MI) was computed simply as $MI = 100\ t/N$, where t is number of persons who have resided in the same house for 5 years or more and N is total population in that census tract.

The mobility index is based upon the average time that residents in a neighborhood occupy a dwelling unit. This indicator does not by itself reveal either negative or positive neighborhood social values. High mobility so defined may increase community cohesion as well as lower housing property values. The effects depend in large part upon the nature of the neighborhood and the socioeconomic characteristics of the in- and outmigrants being studied. In addition, the fact that freeway construction through a neighborhood with a high mobility index may in fact increase the mobility of the neighborhood and the destrictive effect may very well be offset by its positive contribution to labor mobility. Furthermore, the disrupted neighborhood cohesion might not be due as much to the freeway, once constructed, as it is to the changes in the perception of neighborhood identity, street environment changes, residential mix, development characteristics, etc.

A Neighborhood Social Interaction Index (NSII) has been developed to show neighborhood behavior (neighboring, use of local facilities, and participation) and neighborhood perception (identification, commitment, and evaluation). The index can be estimated by using residential mobility (M), percent of residential land (R), and housing units per acre (HU). Mobility has been found to be so important that it alone can be used to provide rough estimates of social interaction changes that might be associated with highways.

Burkhardt [1,2] used the above mentioned three descriptors with the data for West Philadelphia, estimated the functional relationship between NSII and the descriptors and found the equation

$$NSII = 76.29 - 1.45\ M - 0.36\ R - 0.30\ HU$$

has very high coefficient of determination, $R^2 = 0.91$. In recognition of the external effect, Burkhardt finally added to his model another

variable--intraneighborhood accessibility (A). The overall linear model measuring the change of social interaction looks as:

$$NSII = f(-M, -R, -HU, +A)$$

As Burkhardt pointed out, his NSII equation depends vitally upon the mobility variable which in essence is similar to the mobility index described previously. Our first criticism of the mobility index is also applicable to the NSII. However, the NSII may represent an improvement over the mobility index because the indicator has included both positive and negative factors, however subjective they may be, that the lower the NSII, the less disruptive neighborhood effect the highway construction has. Nevertheless, the weights of the four independent variables and their functional relationship with the dependent variable seem to be unduly dominated by the mobility index, and yet its negative impact on social interaction is not well specified and demonstrated, and far from being generally accepted.

The social feasibility model stresses the importance of pedestrian dependency and uses housing and population characteristics to discern and estimate this dependency. Several of the factors beyond walking were also used in estimating pedestrian dependency, e.g., ethnic groups and population age. Thus, pedestrian dependency as used in the social feasibility model to some extent serves as a surrogate for other neighborhood characteristics (such as neighboring). Pedestrian dependency can be calculated for a census tract, a city, or other area. It includes some combination of general pedestrian dependency, school pedestrian dependency, local shopping pedestrian dependency, and social institution pedestrian dependency.

Kaplan, Gan and Kahn [4] found that among the four activity patterns under study, school, shopping and social institutions are significant and important neighborhood-based activities. These activity patterns were therefore incorporated in their social feasibility analysis.

Although criticism can be levied against the social feasibility model (SFM) regarding the selection of variables, this model seems to be better than the mobility index and the neighborhood social interaction index models in that it takes into account a set of social variables concerning the physical environment, human behavior, and economic conditions. Moreover, a rank-order system was developed in the SFM to provide information for setting priorities and choices among alternatives. Its technique resembles the utility and preference ranking of the so-called "marginal analysis" in economics.

Our major criticism of SFM is related to its index structures. First, no theoretical foundation was given to support any of the formulas used. Second, there was no explanation as to why the three variables used should be weighted equally when constructing the index. Third, U.S. median income may be a better variable than city median income for the purpose of standardization. Fourth, would any other form of index construction be more meaningful than the product itself? Finally, why do neighborhoods with high proportions of children or the elderly be overemphasized and treated differently from others?

In short, all models described previously tend to fall short of theoretical foundation and methodological soundness in impact assessment in general and in social welfare evaluation in particular. Neither of these models possesses every basic characteristic essential to a social indicator utility and performance evaluation proposed by Liu [5,6].

The validity of the predictive models delineated in the preceding section were tested empirically by using 1960 and 1970 data from 24 study areas and 21 control areas selected from the four metropolitan areas having circumferential highways--Kansas City, Indianapolis, Omaha and St. Louis.

The principal criteria for selecting the study areas are: (1) the study area must have a new highway that opened up during the 1960's; (2) the census tract is used as a basic unit for impact assessment because it offers the most readily available socioeconomic data required in this study; (3) the selected census tract had a population between 2,500 and 10,000 in 1960; (4) within the population size range, at least one tract each is selected to represent the small, medium, and large neighborhood under study.

The principal criteria for selecting the control areas are the homogenity considerations in: (1) residential and commercial composition similar to the study area; (2) demographic characteristics by size of population similar to the study area; (3) socioeconomic characteristics by medium family income similar to the study area; (4) no freeway passing the area and also somewhat remote from the new highway being studied.

The mobility indicator approach implies that the mobility indicators should be greater in the study area in which a highway segments, than in a control area, i.e., the higher the mobility, the less the description of the highway construction would be. To test this hypothesis, the level of changes in the mobility index between the study and control areas are compared. The results obtained for the four selected metropolitan areas are neither consistent nor conclusive.

The social feasibility model was also tested by calculating the dependence rates for the four cities. According to this school, the higher pedestrian dependence is on walking, and hence, the

more disruptive a highway would be. Therefore, the level of the change in the school pedestrian dependency rates in the study area should be smaller relative to those without a highway in the control area. However, empirical results show that differences in this rate are mostly inconsistent with the underlying hypothesis: the higher the rate, the more vulnerable the neighborhood is to disruption by a highway.

Similar inconsistent patterns emerged in the percentage changes of the local shopping facility pedestrian dependency and the social institutions' pedestrian dependency rates for the study areas in the selected four cities. For the local shopping facility pedestrian rate, of the six study areas in each city, three in Indianapolis, two in Kansas City, four in Omaha and three in St. Louis experienced "unexpected" difference in the rates relative to the control areas. For the social institutions' pedestrian dependency rate, "unexpected" changes occurred in four study areas in Indianapolis, five in Kansas City, four in Omaha and two in St. Louis.

These inconsistent patterns of the changes of both the component and composite pedestrian dependency rates for the four selected cities indicate that the social feasibility model is not an appropriate model for accurately predicting the impact of highway construction on a neighborhood. Stein [11] has recently provided detailed analysis and evaluation on these models.

III. A Neighborhood Quality of Life Production Model

The overall impact of highways should not only be studied for the benefits and costs to the highway users or even the neighborhood's residents, but also should be examined from the nonuser's point of view. In other words, the feasibility of a public investment should be analyzed from the viewpoint of the quality of life of all individuals affected by the investment, directly and indirectly. And if not all user and nonuser benefits and costs are to be studied, the impacts on the quality of life of the neighborhood residents before and after the investment should at least be investigated. A neighborhood impact model was thus recently designed to detect the changes in the quality of life of the neighborhoods in which new highways are constructed and used by the author [7].

For any individual, QOL expresses that set of "wants"--physical (PH) and psychological (PS)--when taken together, that makes the individual happy or satisfied. The concept of quality of life varies not only from person to person, but also from place to place and from time to time. Since most psychological inputs to our Quality of Life are not quantifiable, an empirical measure of the level of quality of life people enjoy must hold the psychological attributes constant, i.e.,

$$QOL_{jt} = f(PH_{jt} \mid PS_{jt})$$

The physical part of the neighborhood quality of life model was then described by Liu [1] as follows:

$$QOL^{s}_{jt} = g[EC(H, EX), ED(H, EX), SE(H, EX), MA(H, EX)]$$

$$QOL^{c}_{jt} = h[EC(EX), ED(EX), SE(EX), MA(EX)]$$

where H denotes highway construction and EX represents all exogenous changes other than highway; the subscripts j and t denote the jth neighborhood and time period t, and the superscripts s and c denote the study and control areas. The variables EC, ED, SE, and MA stand, respectively, the economic, education, social and environmental, and mobility and accessibility components.

The effect of highway construction and other concomitant exogenous changes on the neighborhood's quality of life can be described by:

$$dQOL^{s} = \frac{\partial g}{\partial EC}\left(\frac{\partial EC}{\partial H}dH + \frac{\partial EC}{\partial EX}dEX\right) + \frac{\partial g}{\partial ED}\left(\frac{\partial ED}{\partial H}dH + \frac{\partial ED}{\partial EX}dEX\right)$$

$$+ \frac{\partial g}{\partial SE}\left(\frac{\partial SE}{\partial H}_{H} + \frac{\partial SE}{\partial EX}dEX\right) + \frac{\partial g}{\partial MA}\left(\frac{\partial MA}{\partial H}dH + \frac{\partial MA}{\partial EX}dEX\right)$$

Note that the signs of the partial derivatives of QOL with respect to the four components are all positive, while the signs of the partial derivatives of the four components with respect to H and EX are ambiguous a priori and should be determined via empirical estimation. In the case of control areas where no highway was built, the first term in each of the four brackets on the right-hand side of the least equation vanishes. Thus,

$$dQOL^{c} = \left(\frac{\partial g}{\partial EC}\frac{dEC}{dEX} + \frac{\partial g}{\partial ED}\frac{dED}{dEX} + \frac{\partial g}{\partial SE}\frac{dSE}{dEX} + \frac{\partial g}{\partial MA}\frac{dMA}{DEX}\right)dEX$$

The quantitative effects of highway construction on a neighborhood's physical quality of life may be additively measured and compared by comparing the magnitudes of $dQOL^{s}$ and $dQOL^{c}$. Specifically, if $dQOL^{s}$ is greater (or smaller) than $dQOL^{c}$, then highway construction is likely to be conductive (detrimental) to the physical quality of life of a neighborhood.

More than 30 factors were originally selected to represent the four quality of life components most affected by the highway construction, i.e., economic, education, social and environmental, and mobility and accessibility. The factors were selected on the basis of five criteria: commonality,

simplicity, adaptability, neutrality, and utility [5]. However, due to data problems only 21 variables were practically employed in the model for final impact assessment. Appendix A presents the variables selected and the expected individual variable effect in the four objective components of our quality of life production model. Theoretically the four components are assumed to be independent of each other, and the quality of life level should be viewed strictly as a stock variable--it reflects the degree of human satisfaction at a particular point in time, given the quantity of quality inputs they possess. Practically, some of the assumptions have to be relaxed, e.g., the quality of life output is usually defined over a period of time and hence is a flow variable. Since the factors of both flow and stock variables are relevant for evaluating social well-being, the actual calculation of quality of life indicators involves variables characterized by either stock or flow attributes. Furthermore, the quality of life model developed on the individual basis is also personalized to describe the entire neighborhood on the assumption that individuals in the neighborhood are more or less homogeneous in socioeconomic background and utility considerations.

IV. Neighborhood Impact of Highway Construction: Some New Evidence

The model employed here is in an additive, linear form, and raw data on each individual variable were first standardized and transformed into the conventional "Z" scores such that the mean of the Z scores becomes "0" and its standard deviation becomes "1.0." The basic reason for this standardization is to eliminate the units of measurement among different variables so that they can be neutral and further operated depending only on the direction of those variables toward the explanation of the variations in the quality of life.

An equal weighting scheme was applied to the variables at the same level--subcategory, indicator category, and quality of life component--for simplification sake and future methodological departure as well. In order to avoid the influence of any variable taking on extreme value under such an equal weighting scheme, all "Z" scores were also converted into an ordinal point scale ranging from "1" to "5" based on their percentile distribution with the lowest 20.0 percentile being assigned "1," and the next "2," etc.

Data for all variables listed in Appendix A were collected for the 24 study and 21 control census tracts, earlier mentioned for 1960 and 1970 for the four SMSA's. The composite quality of life indicators were also computed according to the methodology above delineated. Although the changes in quality of life indicators from 1960 to 1970 in both study and control neighborhoods are important, and they do provide us the essential information on the general welfare in each of the neighborhoods over a period of 10 years, it should be noted that the associated changes per se convey no message as to the net effects of a highway on any neighborhood's general welfare. The net effects of a highway may only be reflected through the comparisons of the associated changes 1960 to 1970 between the study and the control neighborhoods. Specifically, if the associated changes for the period are greater (smaller) in the study areas than the counterparts in the control areas, one may conclude that highway construction does have some positive (negative) effects on neighborhood quality of life. In other words, the effects are judged by the ratio of quality of life indicators in the study areas to that in the control areas (S/C)i over the 10-year period. The empirical results for the selected six pairs of neighborhoods in the four metropolitan areas for the quality of life component and overall quality of life indicators are shown in Table 1.

As the results in Table 1 show, when all six pairs of ratios were averaged, nearly all of the four quality of life components received a value greater than unity, except for the economic component in Omaha. This indicates that on the whole highway construction has brought about positive effects on neighborhood life quality on a regional basis, despite the fact that many neighborhood pairs of indicator ratios are less than unity. For example, highway construction had rather negative impacts on socioenvironmental considerations in Indianapolis since four of the six neighborhood pairs showed a ratio value smaller than 1.0 where study areas were compared to the control areas. Similarly, the unfavorable results were shown economically for Omaha and the negative impact was such that it even surfaced to appear at the metropolitan level as shown in the last column of Table 1. Nevertheless, the results, however tentative they are, may still lead one to conclude that, on the average, the construction of a highway has improved neighborhood quality of life about 3.0 percent in Indianapolis and St. Louis, 4.0 percent in Omaha, and 6.0 percent in Kansas City.

It should also be pointed out that the last column in Table 1 represents the major findings of this study. It is conceivable to have lower quality of life indicators in the study neighborhood areas than in the control areas because there are many factors other than highway construction which could affect neighborhood quality of life, i.e., the ratios of (S/C)i could possibly be smaller than unity in some neighborhood areas even though our null hypothesis is that, in general, highway construction enriches neighborhood quality of life. However, the figures in the last column do point out the positive contribution of highway construction to neighborhood quality of life for the metropolitan area as a whole.

Given that there are differences in the metropolitan average comparison of study versus control areas, i.e., the ratios are greater than unity, one would question whether the differences are statistically significant. In other words, are the positive effects so identified for the study areas really different from those for the control areas, and are they statistically different at all from a no-effect nul hypothesis? A simplified Student "t" test suggested by Sandler [10] was performed on the basis of information shown in the last column of the table. The computed "A" statistics for the QOL component indicators is 0.173 and for the QOL indices, it is 0.273. Both of them are smaller than the corresponding critical values of 0.266 and 0.324 at the 5 percent significance level for 23 and 3 degrees of freedom, respectively. Thus, the null hypothesis that the mean QOL values for both control and study areas are equal is rejected. Consequently, the percentage gains in average QOL indicators shown in the last column of the tables mentioned are statistically sustained.

V. Concluding Remarks

Several predictive models of highway impacts on neighborhood, including the mobility index and the social feasibility models, were tested with the data collected from 24 study and 21 control census tracts in the four selected metropolitan areas between 1960 and 1970--Indianapolis, Kansas City, Omaha and St. Louis. Although the usefulness of these models was questioned theoretically, empirical problems of these models did also surface when they were applied to the selected areas for highway impact assessment. In view of the inconsistent and confusing results obtained, the empirical testings seemed to fail to lend support to the validity and the applicability of these predictive neighborhood impact models.

A transport-variant neighborhood quality of life production model was developed with the focus being on the effect of highway construction. The model essentially consists of two QOL production functions expressing the changes in the QOL, respectively, of the study and control areas, in response to the changes in the component indicators as a result of highway construction and other exogenous changes. The effect of highway construction on a neighborhood's quality of life is estimated by summing the effects of highway construction on the transport-related factors which form the basis for the computation of the four QOL component indicators, i.e., economic, education, social and environmental, and mobility and accessibility indicators, and then comparing them to the QOL indicators generated simultaneously for the control areas where no new highways were opened up during the study period. Specifically, the net impacts of highway are to be measured by differential rate of changes between the study areas and the control areas, i.e.,

$$(dQOL^{s}_{jt} \;/\; dQOL^{c}_{jt}).$$

The major findings of the recommended QOL models are that it is indicative, specific and capable of evaluating the construction impacts quantitatively for both purposes of ex-ante prediction and ex-post assessment. The opening-up of highways in the four metropolitan areas did improve the life quality of the affected neighborhoods in numerous accounts including enhanced economic vitality, greater mobility and better accessibility, higher educational attainment, and enriched socio-environmental conditions. For the overall life quality consisting of these four basic components, the results show that a gain of some 3.0 to 6.0 percentage points could be attributed to highway construction. Nevertheless, these are tentative and incomplete results not only because some important variables such as crime rates, property values, noise and air pollution were excluded due to unavailable data but also because the model only attempts to quantitatively measure the physical inputs to our quality of life while holding constant the psychological inputs. Furthermore, it is necessary that the utility of the QOL model and its technical approach be generalized and confirmed with more empirical applications.

REFERENCES

1. Burkhardt, E., "Impact of Highways on Urban Neighborhood, A Model of Social Change," Highway Research Record, No. 356, pp. 85-94 (1971).
2. Burkhardt, J., and J. Rothenberg, Changes in Neighborhood Social Interaction (Washington, D.C.: Federal Highway Administration, 1971.)
3. Hill, S. L., The Effect of Freeways on Neighborhoods (Sacramento: California Division of Highways, June 1967).
4. Kaplan, M., et al., "Social Characteristics of Neighborhoods as Indicators of the Effects of Highway Improvement" (Springfield, VA: National Technical Information Company; 1972).
5. Liu, Ben-chieh, "The Utility of Quality of Life Indicators in Regional Public Decisionmaking," paper presented at the National Conference of the American Association of Public Administration, Washington, D.C., April 19-22, 1976.
6. Liu, Ben-chieh, "A Quality of Life Production Model for Project Impact Assessment," in K. Finsterbush and C. P. Wolf (eds.), Methodology of Social Impact Assessment (Pennsylvania: Dowden, Hutchinson and Ross, 1977).
7. Liu, Ben-chieh, "Models of Highway Construction Impact Assessment," Proceedings of Institute of Environmental Sciences (Mt. Prospect, ILL: The Institute of Environmental Sciences, 1977).
8. McLean, E. L. and W. G. Adkins, "Freeway Effects on Residential Mobility in Metropolitan Neighborhoods," in Highway Research Board, Highway Research Record, No. 356 (1971).

9. McLean, E. L., et al., "Further Investigation of the Mobility Index for use in Predictive Freeway Effects on Neighborhood Stability," Texas Transportation Institute, Texas A&M University College Station, Bulletin 41, June 1970.
10. Sandler, J., "A Test of the Significance of the Difference Between the Means of Correlated Measures, Based on a Simplification of Student's t," British Journal of Psychology, Vol. 46, pp. 225-226 (1955).
11. Stein, M., "Social Impact Assessment Technique and Thin Application to Transportation Decisions," Traffic Quarterly, pp. 297-316, April 1977.
12. Wachs, M., "Social, Economic and Environmental Impacts of Transportation, Systems Resource Paper," in Urban Travel Demand Forecasting, Special Report 143 (Highway Research Board, 1973).

TABLE 1

RATIOS OF QUALITY OF LIFE INDICATORS BETWEEN STUDY AND CONTROL AREAS, 1960-1970

SMSA and QOL Component	Neighborhood Pairs (S/C)1	(S/C)2	(S/C)3	(S/C)4	(S/C)5	(S/C)6	Metro. Av.
Indianapolis							
(EC)	1.06	1.27	1.02	0.72	1.05	1.13	1.04
(MA)	1.20	1.29	1.33	1.15	0.43	0.91	1.05
(Ed)	1.05	1.42	1.23	0.61	1.79	0.56	1.11
(SE)	0.87	0.88	1.79	0.65	0.95	1.47	1.10
Overall	1.02	1.21	1.31	0.78	0.88	0.98	1.03
Kansas							
(EC)	1.33	1.00	0.99	1.31	0.78	0.87	1.05
(MA)	2.66	2.66	0.86	1.05	1.00	0.48	1.45
(Ed)	0.67	1.19	1.57	0.61	0.99	1.08	1.02
(SE)	1.23	0.75	0.96	0.88	1.02	1.19	1.01
Overall	1.36	1.24	1.05	0.94	0.93	0.86	1.06
Omaha							
(EC)	0.65	0.92	1.15	1.05	0.85	1.25	0.98
(MA)	1.17	2.10	1.99	1.03	0.80	0.74	1.31
(Ed)	1.14	1.08	0.92	1.00	0.94	1.00	1.01
(SE)	0.49	1.04	1.16	1.42	1.13	1.32	1.09
Overall	0.87	1.14	1.24	1.10	0.92	1.05	1.04
St. Louis							
(EC)	0.54	1.31	1.04	0.96	1.01	1.19	1.01
(MA)	0.65	1.11	0.43	0.88	1.00	2.00	1.01
(Ed)	0.17	1.26	1.51	1.14	1.09	1.99	1.19
(SE)	1.00	1.44	0.96	0.91	0.94	1.01	1.04
Overall	0.52	1.27	0.91	0.96	1.03	1.49	1.03

SMSA stands for Standard Metropolitan Statistical Area--one or more contiguous counties with a central city having 50,000 or more people.

* The research underlying this paper was supported by a contract (DOT-FH-11-8788) from the Federal Highway Administration to Midwest Research Institute. The helpful assistance and comments of Floyd Thiel and Roger Mingo of FHwA, Eden Siu-hung Yu of Oklahoma University and Mary Kies and Barry Sanders of MRI are acknowledged. The views expressed in this paper are those of the author and he is solely responsible for any remaining shortcomings.

APPENDIX A

NEIGHBORHOOD LIFE QUALITY COMPONENTS AND FACTOR EFFECTS

	Factor Effect
Economic Component	
I. Individual Economic Well-Being	
A. Median family income	+
B. Wealth	
1. Percent of owner-occupied housing units	+
2. Percent of households with no automobiles available	-
3. Median value of owner-occupied single-family housing units	+
II. Community Economic Health	
A. Percent of families with income below poverty level	-
B. Percent of families with income below poverty level or greater than $15,000	-
C. Unemployment rate	-
x D. Land value	
1. Commercial and industrial	+
2. Undeveloped	+
Education Component	
I. Median School Years Completed by Persons 25 Years Old And Over	+
II. Percent of Persons 25 Years Old and Over Who Completed 4 Years of High School or More	+
III. Percent of Persons 25 Years Old and Over Who Completed 4 Years of College or More	+
IV. Percent of Population Ages 3 to 34 Enrolled in Schools	+
x V. Changes in the Elementary School Attendance Rate	+
Social and Environmental Component	
I. Individual Conditions	
A. Existing opportunity for self-support	
1. Labor force participation rate	+
2. Unemployment rate	-
B. Percent of workers working in their county of residence	+
II. Community Living Conditions	
A. Percent of families with income below poverty level	-
B. Percent of housing units lacking some or all plumbing facilities	-
C. Percent of occupied housing units with 1.01 or more persons per room	-
D. Percent of workers using public transportation	+
x E. Acres of parks and recreation areas per 1,000 population	+
x F. Crime rate	-
x G. Population density	-
Mobility and Accessibility Component	
I. Mobility	
A. Percent of persons who have resided in same house for 5 years	-
B. Percent of households with no automobiles available	-
C. Percent of time saved in traveling to city hall	+
x D. Housing segregation index	-
x II. Accessibility	
A. Number of retail establishments built since 1960 (per 1,000 population)	+
B. Number of gas stations built since 1960 (per 1,000 population)	+
C. Hospitals built since 1960 (per 1,000 population)	+
D. Schools built since 1960 (per 1,000 population)	+
E. Parks and recreational areas developed since 1960 (per 1,000 population)	+
F. New housing starts (per 1,000 population)	+
G. Property crime rates (per 1,000 population)	-
H. Traffic count in the busiest intersection in the tract	-

Factors and component marked with x were not included in the study due to data deficiency.

STATISTICS AND MODELING IN URBAN SOCIAL PROBLEMS: DISCUSSION

Frances M. Lerner, University of Baltimore

The traditional definition of a social problem is: "....a condition affecting a significant number of people in ways considered undesirable, about which it is felt that something can be done through collective social action". */ Social problems emerge from changes in values and behavior. Thus a change in values "creates" social problems when conditions once considered either good or bad but inevitable, part of the natural order of things, later are considered bad and changeable. Thus discrimination against minorities in this country, even though long existing, was suddenly defined as bad and not inevitable in the 1960s by significant segments of the electorate. Segregation, the consequence of discrimination, was defined as bad, integration as good. The ramifications were felt in education, housing, employment, public accomodations, voting, and other fields.

Changes in behavior may also "create" social problems, for example when they cause structures to be thrown out of balance. Thus outmigration of a significant number of residents, either from a particular neighborhood or from a city itself to suburbs or elsewhere, creates pockets of social disorganization, leaves cities underfinanced, and breaks up communities.

In American life today urban social problems are highly salient, targets for immediate attack by "collective social action". The papers in this session illuminate a wide range of these problems: busing and school integration, discrimination, white-nonwhite housing differentials, white flight and central city population loss, neighborhood transition, and impact of highway construction on urban neighborhoods. Each paper applies sophisticated techniques to develop statistical measures and/or models of various aspects of these problems.

Casterline**/

This analysis of the demographic correlates of attitudes toward school busing and "integration uses data from the 1976 Detroit Area Study, a probability sample of the entire Detroit SMSA, with overall response rate 75.4 percent. An unusually large sample of Blacks—400 of 1134 respondents—permitted separate analyses for each race, a unique feature of studies of this type. The analysis focusses on race, education, age, and having children in the public schools in relation to two questions, one whether the respondent approved or disapproved of busing to integrate Detroit schools, the second (essentially a school integration question) whether the respondent would object to sending his or her children to a school where more than one-half the children are of the opposite race. The log-linear analysis used in the paper required that all variables, except the busing question, be dichotomized; that question was trichotomized into "approve", "disapprove" and "strongly disapprove". Log-linear analysis permits explicit testing of interactions among variables, as well as of the direct effects of independent variables on the dependent variable. The log-linear models were fitted to the multivariate tables by an iterative proportional fitting procedure as implemented by the ECTA computer program. Only about 7 percent of the Whites approved of busing, in contrast to 50 percent of Blacks, while only 33 percent of Whites and 84 percent of Blacks had no objection to sending their children to schools where more than one-half are of the opposite race. The opposition to busing is thus extreme among Whites, while race differences in responses are large.

The commonly observed relationships of demographic variables with racial attitudes appeared in the analysis of integration but not busing. In the former, both age and race directly affected responses, while the following interactions were also significant: education and race, education and age, and age and race. Age was inversely related to not objecting to sending children to a school with more than half of the children of the opposite race.

Opposition to busing, however, does not follow usual demographic patterns. Race is the primary variable here; among Whites, opposition to busing is spread throughout. The implication, according to the author, is that the career of attitudes on busing may follow a different pattern than previously observed on other racial matters. That is, in the latter gradual public acceptance of more liberal policy has followed initial opposition, especially because usually the vanguard groups have been upper socio-economic strata, especially the better educated. This is not the case here; no vanguard groups exist among Detroit Whites.

A major contribution here is the "new" variable, having children in the public schools. These persons are closest to direct involvement in the busing process; they should be most affected by it, most opposed to busing, but this is not true. They are not more opposed to busing than others, and race is still the basic variable.

The author's expectation (based on surveys), was that there was an overall value change, from acceptance of segregated schooling to integration, and the study tested it. But we really had no adequate explanation of this process of turning desegregation into integration, involving only passive acceptance of "unfair" legislation permitting segregation into active involvement in achieving integration, and it was perhaps too much to expect this. Perceptive observers realized that simply eliminating that form of busing which moved children of both races to segregated schools would not integrate schools where housing patterns, especially in northern cities, had created elementary and junior high schools at least as segregated as the Jim Crow schools of the Deep South.

Relatively few Northern schools were legally segregated (they were de facto), and the only quick

*/ Paul B. Horton and Gerald R. Leslie, The Sociology of Social Problems. Englewood Cliffs, N.J.: Prentice-Hall, 1974. Fifth Edition, p. 4.

**/ John B. Casterline, Demographic Correlates of Attitudes Toward Busing and School Integration.

way to integrate was, in the short run to re-draw boundaries of school districts across neighborhoods and perhaps legal jurisdictions, and in the long run to build new schools in neighborhoods to offset housing segregation. The difficulty here was that busing would be introduced into areas without it previously. This contrasts with rural and Southern areas, where almost all schools required some busing, but with destinations to segregated schools. What is required here, to include the above in the analysis, is comparison of attitudes toward busing in relation to geography (at least Northern vs. Southern, metropolitan versus non-metropolitan residence) age, and previous experience with busing, etc. In addition, degree of satisfaction with school bureaucracy might be included on the basis that some anti-busing attitudes might be explained as really anti-authority; this could be tested by including other questions on busing.

During the first Nixon administration (1968-72), "busing" became the new code word, or symbol, to express sentiment against integration among those not prepared to accept it. This was buttressed by national policy expressed by the White House and Congress (not the Courts), and it raises the question about whether basic underlying values (rather than attitudes toward a specific procedure or mechanism) ever really changed. It highlights the need to study the process and mechanisms of value shifts under the impact of contrary legislation and other symbolic expression from those in power.

This excellent study should be replicated elsewhere in settings where busing met with better or worse fates. Why is it that some areas desegrated with minimum difficulty, others with continued violence?

Frey */

This paper analyzes white flight and central city loss. It uses an "analytic migration framework" to assess the aggregate impact of selected community-level factors on white population losses in central cities of large metropolitan areas. More specifically, it measures the influence of the size of a city's Black population on aggregate white population loss due to the suburban relocation of intra-metropolitan movers and in-migrants to the metropolitan area. The framework separates analytically distinct components of local and long-distance migration streams contributing directly to central city population change. Data were from the 1970 Census, relating to 1965 and 1970 residence.

The community-level factors (racial and non-racial attributes which were the most important determinants of white city-to-suburb movement in an earlier study by the same author) include: percent city Black; city share of SMSA population; suburb-city educational expenditures per capita; suburb-city tax revenues per capita; city crime rate; postwar suburban development; percent of city workers commuting to a suburb; age of central city (interval between city reaching 50,000 and the year 1970); location of city in the Southern Region (as defined by Census Bureau); and an interaction term for percent city Black and location in Southern Region. The last attribute was included because the previous study had suggested that the "white flight" impact of a city's racial composition was most pronounced in non-Southern SMSAs.

The data indicate that percent city Black increases the suburb propensity of city movers, decreases the city propensity of suburb movers and SMSA in-migrants. But none of these effects was great and each was greatly moderated in Southern cities.

The aggregate impact on white city loss attributable to each city's Black population size was assessed in greater detail in three SMSAs: Cleveland, Dayton, and Dallas. Here also increase in percent city Black was associated with net decrease in white population, yet the impact from large differences in number of city Blacks was not substantial anywhere, extremely small in Dallas. In terms of stream-specific components of white city loss, in Cleveland and Dayton racial influences on the destination choices of white SMSA in-migrants contributed to greater city losses than they did on white intrametropolitan movers. This was not true in Dallas, where the impact of race was small on all stream-specific components of population change. Also, in all three cities racial influences had small impact on the destination-choices of suburb-origin movers.

If the finding is correct that percent city Black has only minimal effects on white city flight, clearly we need to seek explanations elsewhere. The "flight" began in the 1930s, preceding the current perception of the problem, i.e., flight from Blacks, school desegregation, crime, etc.. We should return the study of migration to focus on the changing nature of the city in modern life. By 1930 the older in-migrant flow (rural Americans and, Europeans) to older American cities was to some degree matched by a steady flow of urban dwellers to the suburbs. After World War II the process exploded due to a whole series of technological and administrative changes (cars and highways, FHA and Veteran mortgages, etc.).

Thus improvements in transportation and communication permitted decentralization, not only of residence, but also of business and financial, artistic, and creative activities formerly monopolized in the city. Megalopolis, a term coined years ago, became a sociological reality desipte its political unacceptability. Frey's study, and additional studies along these lines, should redirect analysis of urban problems beyond the themes of White flight from the Blacks and the common bewailing of the disintegration of city life.

El-Attar, Rubin, and Al-Maryati*/

The third paper investigates White non-White housing quality differentials in 1970, holding

*/William H. Frey, White Flight and Central City Loss: Application of an Analytic Migration Framework.

*/M.E. El-Attar, R.M. Rubin, and M.S. Al-Marayati, White-Nonwhite Housing Quality Differentials in the United States: 1970.

income constant. Data were from the U.S. 1970 Census of Housing. The data showed a direct relationship between income level and housing quality for both owners and renters, Blacks and Whites. Also, on average Blacks occupied poorer housing units than Whites.

A two-way analysis of variance tested the null hypothesis of no difference in quality of units rented or owned by Blacks and Whites at the same income. The hypothesis was rejected in three cases (good quality, owner; good quality, renter; and poor quality, owner), and accepted for poor quality, renter.

To assess the source of variation (and explain the above relationships) an analysis of variance for contrasts indicated significant differences in housing quality between Blacks and Whites for plumbing in rented units where incomes were under $7,000 and for plumbing and room density in owner-occupied units at $7,000 and over.

Students of social change are unlikely to be surprised by these conclusions. It is another example of the lag between policy (pro-integration during the 1960s) and social patterns.

But simply to say that objectives were not completely achieved by 1970 is inadequate; we need to know how much change did occur, so that the analysis should also have been done on 1960 data. However, there are some difficulties here; the one table of changes between 1950 and 1970 indicates that the definition of "sound quality" may no longer apply since plumbing is now ubiquitous in urban areas, almost doubling even in rural areas.

Continuous monitoring of housing differentials is indicated to measure effects of changes in values and laws over time. We have no measure of how quickly these changes can be translated into social patterns in a non-totalitarian society.

Cole and Baldus */

The fourth paper deals with statistical modelling to support a claim of covert intentional discrimination against minorities. Conclusions drawn from statistical analyses often form an important and accepted component of the evidence in these cases, but the procedures themselves leave many methodological questions unresolved in a still emerging legal field. The substantive areas involved are employment selection, promotion, school admission, some instances of criminal sentencing, and jury selection. Recently the Supreme Court has ruled that such challenges, on constitutional grounds, require proof of two facts: 1) The existence of discriminatory impact, relative disadvantage accruing to the plaintiff as a result of the suspect practice, and 2) The existence of an intent to discriminate underlying the practice.

Selection processes are classified by the authors in accordance with the amount of discretion left to the decision-maker: a "guided" discretionary process where the decision-maker's choices are influenced in part by qualifications openly stated and measurable, a "purely" discretionary process where no such qualifications are cited. In purely discretionary, five simple measures of the treatment accorded the minority are discussed: 1) The selection or pass rate; 2) The rejection or fail rate; 3) The inverse of the selection rate; 4) The minority representation rate in the post-selection pool; or 5) The actual number of minority candidates chosen.

These measures are examined in measuring discriminatory impact and inferring discriminatory motive. Guided discretionary selection processes are examined more closely, and the overall approach is applied to recent issues.

Statistical procedures can clearly establish the probability of discriminatory behavioral patterns in aggregate behavior, i.e., when representation in a "selection group" of members of a specified population sub-group differs significantly from expected by chance alone with random selection, a discriminatory behavioral pattern can be inferred. However, the danger lies in the imputation of bias in a legal sense, by statistics alone in cases involving single or small numbers of events, i.e., hiring one individual, selection of a single jury, etc.. This crosses levels of conceptualization, from behavioral patterns based on numbers of cases to imputation of discriminatory motive in a single case.

By providing statistical procedures, the authors in effect develop an operational definition of discrimination. The hidden booby trap here is that operational definitions in public life tend to become fixed, rigid, and in the end may confuse rather than clarify. Social science measurements themselves may become social data. This is what happened with the definition of poverty; once fixed, it was no longer responsive to changing social patterns. Sixty-five as retirement age is another example. We run the danger here of operationally defining a concept, and then permitting its operational definition to set norms and often subvert the very process it was initially intended to promote. In discrimination, statistical probability readily translates into quota systems.

Sanathanan, O'Neill, and McDonald */

This paper discusses the fitting of epidemiologic models to panic selling in urban neighborhoods using time-series data on real estate sale prices. Panic selling is seen to resemble an epidemic, peaking at 100 days and then subsiding. The policy implications of this model are discussed, especially with regard to allocation of community development funds among neighborhoods according to future prospects and to identification of neighborhoods needing assistance.

*/James W.L. Cole and David D. Baldus, Statistical Modeling to Support a Claim of Intentional Discrimination.

*/Lalitha Sanathanan, William O'Neill, and John McDonald, Dynamic Modeling of Neighborhood Transition. (This discussion was written on the basis of abstract and oral presentation only.)

Liu */

The final paper evaluated four "impact models" (really measures) of the effect of highway construction on an urban neighborhood. These were applied to 1960 and 1970 Census data for six study neighborhood areas (each a census tract) and almost the same number of control areas in each of four major metropolitan areas: Indianapolis, Kansas City, Omaha, and St. Louis. In addition to being similar to the study areas, each control area was selected because no freeway passed it and it was somewhat remote from the new highway. Application of the first three models gave relatively unsatisfactory values.

The fourth model (the transport-variant quality of life production model) has four components: economic, educational, social and environmental, and mobility and accessibility. More than 30 factors were originally selected to represent these components, but only 21 were used in the model. Net effects were reflected through comparisons of changes 1960 to 1970 for study and control neighborhoods. In the four metropolitan areas, highways improved the quality of life in all major components by 3 to 6 percentage points, except for the economic component in Omaha.

If this paper illustrates anything, it is that indicators of qualitative concepts (quality of life) need constant re-definition simply because social patterns are constantly changing, and that the unanticipated consequences of purposive social action today become the social problems of tomorrow. Highways were originally constructed to improve the quality of life; many feel that they do just the opposite. Next year's session will deal with the impact of subways on some aspect of quality of life, even these are now promoted as panacea for current problems.

*/Ben-Chieh Liu, Impact Models of Highway Construction on Urban Neighborhoods.

EVALUATION OF THE 1972-73 CONSUMER EXPENDITURE SURVEY

Cathryn Dippo, John Coleman and Curtis Jacobs, Bureau of Labor Statistics

The Consumer Expenditure Survey (CES) was a major component of the Bureau of Labor Statistics (BLS) program to update the Consumer Price Index (CPI). Its primary purpose was to collect relative annual mean expenditures for all components of consumption to be used as the basis for creating the cost weights of the revised Consumer Price Index (CPIR). Additionally, the survey provided data for publication of mean expenditures at various geographic and demographic levels.

The CES consisted of two separate types of questionnaires, a diary and an interview, administered to independent samples of housing units. The diary survey was used primarily to obtain data on frequently purchased items; the interview survey for less frequently purchased items.

I. Sample Design and Procedures

The specific PSU (Primary sampling unit) design for the CES was a modified CPS design of 216 PSU's, 30 self-representing (SR) SMSA's and 186 non-self-representing (NSR) PSU's. The original plan was to complete the survey in one year; however, due to a reduction in funds, the data collection was divided into two one-year phases. One half of the selected housing units in SR areas were interviewed each year; all the housing units in half of the NSR PSU's were interviewed each year.

The eligible population was composed of all civilian noninstitutional persons and certain persons residing in group quarters. A systematic unclustered sample of approximately 15,000 housing units was selected for each year of the diary survey. A similar sample of about 13,000 housing units was selected for the interview survey. Each housing unit in the diary sample was requested to complete two one-week diaries and was assigned an initial week of interview so as to distribute data collection over the period July 1972 to June 1974. For the interview survey, the sample housing units were interviewed during the first quarter of 1972 or 1973 and for four succeeding quarters for a total of five interviews per household.

Approximately eleven to thirteen percent of the housing units designated for the interview survey were vacant, non-existent or ineligible; another ten percent refused or were unable to be contacted. For the diary survey, about thirteen to fifteen percent of the units were ineligible. Seventeen percent refused to cooperate during the first year and nine percent the second year. Therefore, diaries were completed at about 10,000 units the first year and 12,000 the second. During the last quarter of the interview survey, about 10,000 units were interviewed each year.

A sampling weight was determined for each consumer unit (CU)[1] responding in the fifth quarter of the interview survey and each consumer unit completing at least one week of the diary. The weight included factors for noninterview adjustment, a ratio adjustment for NSR PSU's by color-residence, a ratio adjustment to population controls by age-sex-race, and a CU adjustment based upon multiple-CU household composition. These procedures provided estimates consistent with the number of households estimated by the March Current Population Survey (CPS).

Data collection and processing of the data to this stage was completed by the Bureau of the Census. After the data tapes were transmitted to BLS, additional processing included editing, allocation, imputation, annualization, and sales tax adjustments. Two separate data bases were created - one for CPIR and another for publication. The results presented in this paper are for the most part based upon the data base developed for the CPIR. Therefore, levels of mean expenditures presented here are given in terms of the CPIR classification scheme and may not agree exactly with those developed by BLS for other purposes. Moreover, data from the diary survey for infrequently purchased items will not be published by BLS and are used in this paper only for analytical purposes.

II. Sampling Errors

Estimates of sampling errors have been simulated by using the random group and collapsed strata methods. Basically, each designated housing unit has been systematically assigned to one of t random groups in the order of selection. The assignment is independent between SR and NSR PSU's, but is across PSU's within type. For the diary, there are 10 random groups; for the quarterly, fifteen. In addition, the SR PSU's are grouped into 15 clusters and the NSR PSU's for the first year into 43 clusters. For both years combined there are 93 NSR clusters. The NSR clusters have been formed by grouping together two or three PSU's of similar size and characteristics.

Using the notation: c = cluster, g = random group, t = number of random groups, X_{cg} = expenditures in cth cluster, gth random group, $X = \sum_c \sum_g X_{cg}$ = total expenditures (either SR or NSR), the random group estimates of variance are given by:

$$\hat{\sigma}^2_X = \frac{1}{(t-1)} \sum_c \{t \sum_g X_{cg}^2 - [\Sigma X_{cg}]^2\}$$

This is an estimate of the within - PSU component of variance and has been computed for both the SR and NSR PSU's (σ^2_{SR} and σ^2_{NSRW}). Although this method of variance estimation tends to slightly overstate the variance, it does include the effects of both the weighting and systematic sampling procedures.

To estimate the total variance for the NSR PSU's, collapsed stratum estimates have been made as follows. Let: i = PSU, c = cluster, k = number of PSU's in cluster, $X_{NSR} = \sum_c^{NSR} \sum_i X_{ci}$ = total NSR expenditure, P_{ci} = proportion population of stratum represented by PSU i is to the total population of the cth cluster, then

$$\hat{\sigma}^2_{X_{(NSR)}} = \Sigma[k\Sigma_i (X_{ci} - X_c P_{ci})^2] .$$

The total variance of expenditures at the U.S. level is then estimated by

$$\hat{\sigma}^2 = \hat{\sigma}^2_{X(SR)} + \hat{\sigma}^2_{X(NSR)}$$

and the between PSU component of variances $\hat{\sigma}^2_B$ by

$$\hat{\sigma}^2_B = \hat{\sigma}^2_{X(NSR)} - \hat{\sigma}^2_{X(NSRW)} \cdot$$

Since the variances of primary concern are those for mean expenditures, a ratio of two random varibles, the variances of consumer units and the covariances between expenditures and consumer units have been computed using similar procedures. The relvariance of the ratio $X = X/Y$ is then estimated by $V^2_{\bar{X}} = V^2_{\bar{X}} + V^2_{\bar{Y}} - 2V_{XY}$.

At the U.S. level, the relative proportions of total variance of consumer unit wks (142,341,000) due to SR, NSRW, and between PSU variance are 14, 29, and 57 percent respectively. The largest component of variance is the between, which is a function of the number of PSU's (93).

The average relative between PSU contribution to the variance of mean expenditures for the twenty food EC's is 30 percent, which is about half of the same proportion for the variance of consumer unit weeks. In other words, the effect of having a relatively small number of PSU's is subordinate to the sample size in determining the variance of mean expenditures.

III. Comparisons of Expenditures Between the Diary and the Interview

Although the purpose of the diary was to obtain expenditure data for food and other frequently purchased items, respondents were requested to enter all purchases including clothing, household textiles, furniture, appliances, etc.

Therefore, one of the research topics has been the comparison of mean expenditure levels for infrequently purchased items between the diary and the interview surveys. The completion of this task has not been straight forward. Numerous definitional differences[2] exist between the two sources, and the coding schemes for the two surveys are not comparable. Some of these problems have been overcome by using the CPIR data base, but others have required massive recoding. Table 1 presents corresponding mean expenditures from the two surveys with their variances as computed from the CPIR data base along with the absolute differences (Δ) and a measure of significance testing (Δ/σ_Δ). The variance of the difference σ^2_Δ (σ^2 Diary + σ^2 Interview) was computed assuming total independence between the two estimates and therefore may be a slight overestimate. Again, it should be noted that the data presented in these tables was prepared for research purposes only and may not correspond exactly to the final BLS published data. [3]

Comparisons between the two sources (i.e., diary and interview) should not be based upon statistical significance alone; for between estimates not significantly different, the one with the lower coefficient of variation (CV) can be considered more reliable. Also, comparisons for EC's 33, 55, and 64 are relatively meaningless since the interview did not cover many of the items in these EC's.

Of the 47 EC's for which comparisons have been made on 1972-73 data, only nine EC's have non-significant differences between the diary and interview means. These are: 23-maintenance and repair services, 25-other fuels, 26-gas and electricity, 29-furniture, 39-girl's apparel, 50-insurance, 56-professional services, 57-hospital and other medical care services and 60-sporting goods and equipment. In all cases the coefficient of variation for the diary estimate is larger than that of the interview.

For the non-food at home EC's (19-68), the CV for the diary is less than that for the interview in the following EC's: 19-food away from home, 20-alcoholic beverages, 33-housekeeping supplies and 64-toilet goods and personal care appliances. The diary has been used as publication source for these expenditures along with those for the following EC's:

EC	Rel Mean Exp	CV_D/CV_Q
25-other fuels	D=Q	1.3
27-other utilities	D>Q	1.7
47-gasoline	D<Q	1.2
63-tobacco	D<Q	1.0
65-personal care serv	D<Q	1.7

However, for integration, the interview survey has been used as the source of mean expenditure for these EC's.

On the other hand, the CPIR cost-weights for the following non-food EC's are based upon the mean expenditures from the diary:

EC	Rel Mean Exp	CV_D/CV_Q
33 [4]	D>Q	.4
47	D<Q	1.2
55 [4]	D>Q	1.2
59 [4] (part)	D>Q	1.2
61 [4] (part)	D>Q	2.3
64 [4]	D>Q	.6

The quarterly has been used for the remaining EC's including 27, which has a higher diary mean expenditure than quarterly. Therefore, for only EC 47 might the diary have been a better source.

IV. Diary reporting by day of week

In the past, diary surveys have exhibited differentials in levels of expenditure reporting between weeks and days within a week. The 1972 diary survey is no different. Table 2 shows mean expenditures by week for the diary published EC's and indicates the relative differences (Δ/σ_Δ). When only completed diaries are considered for the 27 EC's shown, 21 have greater means the first week than the second week based on 95 percent confidence intervals. Over all diaries, 16 EC's have greater means for week one than week two.

Of 1809 CPI items examined, mean expenditures for only 97 differ significantly between week one and week two. Of these 97 items, differences for 58 are associated with the published diary EC's and 55 the EC's used for CPIR. Not all of these differences result from higher first week mean expenditures. Twenty-one of the 97 items have higher second week means. However, among the items in the diary published EC's, there are only three with higher second week means. As to the reliability of first versus second week mean expenditures, the lower CV's are evenly divided between weeks over all items.

An examination of EC mean expenditures by day of reporting over the 14-day period (See Table 3) shows that the mean expenditure for day one is greater than every other day for all EC's except EC 3-Beef, EC 7-Fish and Seafood, and EC 20-Alcoholic beverages. (It should be noted that a one cent difference may be "significant"; however, the variance on the variance of very small mean expenditures could more than account for this and make such tests meaningless.) The differences between the second day of each week are very small - only three are greater than ten cents. Between days seven and fourteen, the last day of each week, only eleven EC's show differences but none of the food EC's differ by more than one cent making the differences not really meaningful. This implies that the last day of week two is not different from the last day of week one.

If the first day is ignored, most differences are small or within sampling error. The difference between day one and the other days could be due to: telescoping, failure to understand or follow instructions, or completion of all or part of the diary by the respondent using recall methods. Diaries completed either totally or partially by the interviewer using recall methods have lower mean expenditures than those completed by the respondent. This is not an unusual phenomenon. Most interviewers entered recall expenditures in day one and, therefore, to the extent the interviewers did not answer the completion code correctly, the day one mean expenditures for "completed" diaries are biased.

If the first day of each week is dropped and the mean expenditures of the remaining six days compared between weeks, there are only a few EC's with significant differences, even these have small differences. Therefore, it seems reasonable to attribute most of the difference between weeks to the same cause(s) as the first day bias.

V. Implications for CCES

The Bureau of Labor Statistics plans to initiate, some time in 1979, a Continuing Consumer Expenditure Survey (CCES). As presently formulated, this will be an ongoing effort consisting of both a diary and an interview survey in independent samples of approximately 4,800 interviewed households per year within the 86 urban CPI PSU's and an additional 16 PSU's selected to represent the rural U.S. population. The interview questionnaire will be modified to correspond closely with the CPIR item structure, and both the diary and the interview will be modified to include some point-of-purchase (POPS) information. Currently, BLS conducts a separate survey to obtain POPS data for use in selecting the outlets for CPI pricing.

Our ultimate goal is to initiate a set of surveys that will provide the data necessary to update both CPI outlets and cost weights as needed, with as much reliability as cost effectively possible. Evaluation of the 1972-73 CES at BLS has been directed towards this end.

Using the sample sizes planned for the CCES and the CES variances, estimates of variance have been projected for the CCES. These indicate BLS should be able to publish from the diary survey quarterly mean expenditures at the EC level for those EC's presently published from the diary. Other EC's not currently published from the diary but which are projected to have CV's within the range of the food EC's are: 54-Prescription drugs and 59-Reading materials. After four years, under relatively stable economic conditions, estimates of mean expenditures for the food EC's could be made at the market basket level. Therefore, at any time after four years, BLS would have the data necessary to update the cost weights of the CPI.

An indicated below, detailed analysis of the relative effects of the decreased sample sizes (5848 SR in 1972 to 2560 in CCES; 4831 NSR in 1972 to 2245 in CCES) and different number of NSR PSU's (93 in 1972 and 74 in CCES) shows the NSR sample size is far more important than the number of NSR PSU's.

For the CCES diary the estimate of mean expenditures ($\bar{X}$) will be of the form $\bar{X} = \sum_h P_h \bar{X}_h$ where h indicates market basket and P_h is the proportion the market basket population is of the total U.S. Remembering the relationships from Section II,

$$V^2_{\bar{X}} = P^2 V^2_{\bar{X}_{SR}} + (1 - P)^2 V^2_{\bar{X}_{NSR}}$$

$$V^2_{\bar{X}_{SR}} = V^2_{X_{SR}} + V^2_{Y_{SR}} - 2V_{XY_{SR}}$$

and $$V^2_{\bar{X}_{NSR}} = V^2_{X_{NSR}} + V^2_{Y_{NSR}} - 2V_{XY_{NSR}}$$

where P is the proportion of total U.S. population in SR PSU's, the relvariance of $\bar{X}$ can be expressed as

$$V^2_{\bar{X}} = \sum_h P^2_h \left[\frac{\bar{V}^2_{SR}}{n_h} + \frac{\bar{V}^2_{NSR}}{n_h} + \frac{\bar{V}^2_B}{L_h}\right]$$

where $\bar{V}^2$ indicates unit relvariance, n_h is the sample size in the hth market basket and L_h is the number of PSU's in the hth market basket.

$$V^2_{\bar{X}} = \bar{V}^2_{SR} \sum_h \frac{P^2_h}{n_h}{}^{SR} + \bar{V}^2_{NSRW} \sum_h \frac{P^2_h}{n_h}{}^{NSR} + \bar{V}^2_B \sum_h \frac{P^2_h}{n_h}{}^{NSR}$$

For the 1972 diary,

$\bar{V}^2_{SR} = 5848\ V^2_{SR}$, $\bar{V}^2_{NSRW} = 4831\ V^2_{NSRW}$,

$\bar{V}^2_B = 93\ V^2_B$, and for the proposed CCES

$$\sum_h \frac{P^2_h}{n_h}{}^{SR} = .00008167, \quad \sum_h \frac{P^2_h}{n_h}{}^{NSR} = .00016478,$$

$\sum_{h}^{NSR} \frac{P_h^2}{L_h} = .00532058$; therefore,

$$V^2_{\bar{X}_{CCES}} = V^2_{SR_{72}} (.4776) + V^2_{NSRW_{72}} (.7961) + V^2_{B_{72}} (.4748)$$

and unless the between PSU relvariance is very large, the within PSU component is far more significant in determining the reliability for CCES.

The diary data has also been examined by week to determine what kind of reliability would be achieved if consumer units were requested to complete a one week or a three week diary. For the 29 currently published EC's, the average increase in unit relvariances for SR and NSRW from using only one week would be 65 and 72 percent, respectively. If there were no correlation between weeks, a two week diary would have the same effect as doubling the sample size and the increase in unit relvariances from using only one week rather than two would be 100 percent instead of 65 or 72 percent. However, the between PSU variance remains about the same so that the projected CV's are only 22 percent higher. Only EC 25 has a projected CV greater than ten percent based upon one week's data.

The question of primary concern is to determine the optimum number of weeks of diary keeping in terms of both cost and reliability. The variance of a mean expenditure from a "w" week diary can be expressed as:

$$\text{Var } \bar{X}_w = \text{var} \left[\frac{\sum_{i=1}^{n} \sum_{j=1}^{w} X_{ij}}{wn_w}\right]$$

$$= \frac{\sigma^2}{(wn_w)^2} (wn_w) + \frac{2\sigma^2 n_w}{(wn_w)^2} [(w-1)\rho_1 + (w-2)\rho_2]$$

$$= \frac{\sigma^2}{(wn_w)} \{1 + \frac{2}{w} [(w-1)\rho_1 + (w-2)\rho_2]\}$$

For equal reliability from a "w" and "w'" week diary (w>w') assuming equal means and equal response levels for each week,

$$\frac{V^2_{\bar{X}_w}}{V^2_{\bar{X}_{w'}}} = \frac{w'n_{w'} \{1 + \frac{2}{w} [(w-1)\rho_1 + (w-2)\rho_2]\}}{wn_w \{1 + \frac{2}{w'} [(w'-1)\rho_1 + (w'-2)\rho_2]\}}$$

or

$$\frac{n_w}{n_{w'}} = \frac{w' \{1 + \frac{2}{w} [(w-1)\rho_1 + (w-2)\rho_2]\}}{w \{1 + \frac{2}{w'} [(w'-1)\rho_1 + (w'-2)\rho_2]\}}$$

The variable cost for the diary operation C_{t_w} consists of an initial cost $C_f n_w$ and a variable cost associated with each completed diary and return visit $C_v w n_w$

$$C_{t_w} = C_f n_w + C_v w n_w$$

For equivalent cost to obtain equal reliability,

$$\frac{C_{tw}}{C_{tw'}} = \frac{C_f n_w + C_v w n_w}{C_f n_w + C_v w' n_{w'}}$$

$$= \frac{(C_f + C_v w)}{(C_f + C_v w')} \frac{w' \{1 + \frac{2}{w} [(w-1)\rho_1 + (w-2)\rho_2]\}}{w \{1 + \frac{2}{w'} [(w'-1)\rho_1 + (w'-2)\rho_2]\}}$$

For the comparison of a three week to a two week diary,

$$\frac{C_{t_3}}{C_{t_2}} = \frac{(C_f + 3C_v)\ 2\{1 + \frac{2}{3}(2\rho_1 + \rho_2)\}}{(C_f + 2C_v)\ 3\{1 + \rho_1\}}$$

For a two week to a one week diary,

$$\frac{C_{t_2}}{C_{t_1}} = \frac{(C_f + 2C_v)\ \{1 + \rho_1\}}{(C_f + C_v)\ 2}$$

Assuming a thirty percent cost differential between first and succeeding visits, fourteen of the 29 diary published EC's have week-to-week correlations (Table 4) large enough to warrant a one week diary. Twelve of these fourteen EC's are food EC's. The smallest correlation among the food EC's is .23 and among the published EC's, -.12 for EC 25-Other home heating fuels. Only five of the 29 EC's have low enough correlations to warrant a three week diary.

For the CPIR, only EC's 55, 61 and 64 have correlations low enough to warrant a three-week diary. As for the published EC's, the number of EC's with week-to-week correlations greater than or less than .39 is about evenly split. The weighted average correlation of CPIR diary EC's is .40, which only indicates a one-week diary if the relative costs of the first visit versus succeeding visits are about equal. Also, as the level of aggregation decreases to item strata and item, the correlations decrease, indicating a two-week diary is probably optimum for CPI needs.

Although the analysis is not complete and indeed, it has barely started with respect to the interview survey, the diary appears to have succeeded in improving the reliability of frequently purchased items. The small number of NSR PSU's does not appear to be the major factor in determining the reliability; however, despite the small sample size it is expected that CCES will provide four year cumulative data for CPI comparable to the 1972-73 survey. A completely definitive statement on the adequacy of the sample for CPI cost weights cannot be made until the effect of the variance of the cost weights on the index can be examined.

1/ A consumer unit is a single financially independent consumer or a family of two or more persons living together, pooling incomes and drawing from a common fund for major expenditures.

2/ The diary does not include expenditures for items purchased while away from home on vacation.

3/ The interview clothing expenditures do not include expenses for items purchased as gifts.

4/ The quarterly does not cover many of the items in these EC's.

TABLE 1 ANNUAL MEAN EXPENDITURES - 1972 - 73

EC		Diary $\bar{X}$	Diary $\sigma\bar{X}$	Diary CV	Quarterly $\bar{X}$	Quarterly $\sigma\bar{X}$	Quarterly CV	CPI	Δ D-Q	$\sigma\Delta$	$\frac{\Delta}{\sigma\Delta}$
19	Food away from home	452.50	6.50	.0144	427.73	6.76	.0158	D	24.77	9.38	2.64
20	Alcoholic beverages	111.26	2.34	.0210	82.49	1.89	.0229	D	28.77	3.01	9.55
21	Pure Rent (Renters)	536.52	12.15	.0226	622.85	9.80	.0157	Q	-86.33	15.61	-5.53
23	Maint. & Repair Service	90.39	9.59	.1061	88.69	2.72	.0307	Q	1.70	9.96	.17
24	Maint. & Repair Comm.	85.63	4.50	.0526	33.06	1.00	.0302	Q	52.57	4.61	11.41
25	Fuels	66.64	2.82	.0423	73.41	2.44	.0332	Q	-6.76	3.73	-1.81
26	Gas & Electricity	272.37	4.39	.0161	274.27	3.42	.0125	Q	-1.90	5.57	-.34
27	Other Utilities & public services	268.12	4.82	.0180	239.23	2.49	.0104	Q	28.89	5.43	5.32
28	Textile house furnishings	66.90	2.35	.0351	54.41	1.05	.0193	Q	12.49	2.58	4.85
29	Furniture	135.22	12.99	.0961	135.41	3.18	.0235	Q	-.20	13.37	-.01
30	Household appliance	84.01	5.89	.0701	100.41	1.63	.0162	Q	-16.40	6.12	-2.68
31	TV, radio & sound equip	71.57	4.85	.0678	103.36	1.64	.0159	Q	-31.79	5.12	-6.21
32	Other household equip.	182.38	6.11	.0335	60.03	1.35	.0225	Q	122.35	6.26	19.54
33	Housekeeping supplies	134.96	1.69	.0125	43.33	1.30	.0300	D	91.63	2.13	43.09
34	Housekeeping services	143.81	4.34	.0302	128.27	2.84	.0221	Q	15.54	5.19	2.99
36	Men's apparel	103.91	3.78	.0364	146.32	1.81	.0124	Q	-42.40	4.19	-10.11
37	Boy's apparel	24.64	1.20	.0487	36.67	.66	.0180	Q	-12.03	1.37	-8.78
38	Women's apparel	193.54	6.17	.0319	213.02	2.77	.0013	Q	-19.49	6.77	-2.88
39	Girl's apparel	41.88	1.76	.0420	44.23	.88	.0199	Q	-2.35	1.97	-1.19
40	Foot wear	93.70	2.32	.0248	81.86	.73	.0089	Q	11.84	2.44	4.86
41	Infants & Toddlers apparel	16.41	.59	.0360	12.92	.35	.0271	Q	3.50	.69	5.07
42	Sewing material & notion	17.11	.55	.0321	20.95	.43	.0205	Q	-3.84	.70	-5.50
43	Jewelry & luggage	38.04	1.91	.0502	44.15	1.55	.0351	Q	-6.11	2.47	-2.48
44	Apparel Services	40.04	1.01	.0252	61.83	.85	.0137	Q	-21.78	1.32	-16.51
45	Purchase of new cars, trucks, etc.	286.03	23.09	.0807	446.81	10.27	.0230	Q	-160.78	25.27	-3.97
46	Purchase of old cars, trucks, etc.	93.59	21.07	.2251	269.63	6.07	.0225	Q	-176.04	21.93	-8.03
47	Gasoline, motor oil coolant, etc.	357.98	4.61	.0129	399.99	4.26	.0107	D	-42.00	6.27	-6.70
48	Parts & equipment	65.48	3.12	.0476	71.98	.93	.0129	Q	-6.50	3.25	-1.99
49	Maintenance & repairs	127.59	4.03	.0316	138.74	1.98	.0143	Q	-11.14	4.49	-2.48
50	Insurance	203.48	7.77	.0382	196.63	1.94	.0099	Q	6.84	8.01	.85
52	Vehicle rental, regis. & fees	43.55	1.91	.0439	64.24	1.19	.0185	Q	-20.69	2.25	-9.19
53	Public transportation	72.00	4.44	.0617	98.61	3.15	.0319	Q	-26.60	5.44	-4.89
55	Non prescription drug & medical supplies	61.58	1.86	.0302	16.50	.43	.0261	D	45.07	1.91	23.64
56	Professional services	198.36	7.22	.0364	188.99	3.40	.0180	Q	9.37	7.98	1.17
57	Hospital & Other medical care services	38.21	5.45	.1426	32.77	2.41	.0735	Q	5.44	5.96	.91
58	Health Insurance	55.40	3.00	.0542	155.07	2.71	.0175	Q	-99.68	4.04	-24.66
59	Reading Materials	59.39	1.03	.0173	43.79	.62	.0142	D,Q	15.60	1.21	12.94
60	Sporting Goods & Equip.	70.68	11.21	.1586	78.82	4.04	.0513	Q	-8.14	11.92	-.68
61	Toys, hobbies & other entertainment	12.24	6.70	.0557	52.32	1.26	.0241	D,Q	67.93	6.82	9.97
62	Admission fees & other entertainment services	197.28	5.83	.0296	167.58	3.66	.0218	Q	29.70	6.89	4.31
63	Tobacco products	111.84	1.50	.0134	125.92	1.61	.0128	Q	-14.08	2.20	-6.39
64	Toilet goods & personal care appli.	76.56	1.07	.0140	8.26	.18	.0218	D	68.30	1.09	62.89
65	Personal care services	68.49	1.38	.0201	91.86	1.08	.0118	Q	-23.37	1.76	-13.31
66	School books & supplies	12.10	1.05	.0868	18.41	.66	.0359	Q	-6.31	1.20	-5.24
67	Tuition & school fees	67.84	5.51	.0812	93.18	2.90	.0311	Q	-25.34	6.23	-4.07
68	Legal, bank, acc'g funeral and other	95.29	19.00	.1994	20.43	.86	.0421	Q	74.86	19.02	3.94

TABLE 2 Comparison of Weekly Mean Expenditures for Completed[1/] Diaries - 1972

EC	$\bar{X}_1$	$\bar{X}_2$	$\Delta=\bar{X}_1-\bar{X}_2$	$\sigma\Delta$	$\Delta/\sigma\Delta$	EC	$\bar{X}_1$	$\bar{X}_2$	$\Delta=\bar{X}_1-\bar{X}_2$	$\sigma\Delta$	$\Delta/\sigma\Delta$
01	.75	.70	.05	.025	2.02*	15	.83	.76	.07	.027	2.56*
02	2.34	2.15	.19	.052	3.67*	16	.69	.62	.07	.022	3.17*
03	3.97	3.64	.33	.178	1.85	17	1.98	1.78	.20	.054	3.67*
04	2.23	2.07	.16	.069	2.32*	18	2.21	2.08	.13	.050	2.61*
05	1.16	1.09	.07	.039	1.81	19	9.26	8.93	.33	.277	1.19
06	1.08	.96	.12	.038	3.16*	20	2.35	2.17	.18	.110	1.63
07	.72	.67	.05	.039	1.30	27	6.14	5.25	.89	.275	3.24*
08	.60	.53	.07	.017	4.07*	33	3.14	2.84	.30	.096	3.12*
09	2.03	1.90	.13	.052	2.50*	47	7.64	6.94	.70	.218	3.21*
10	1.50	1.37	.13	.039	3.37*	55	1.34	1.32	.02	.099	.20
11	1.02	.88	.14	.046	3.04*	63	2.36	2.21	.15	.070	2.15*
12	1.12	1.02	.10	.031	3.21*	64	1.85	1.67	.18	.061	2.94*
13	.83	.75	.08	.029	2.79*	65	1.60	1.43	.17	.068	2.49*
14	.89	.81	.08	.026	3.07*						

*Significant difference (95%)

1/ Diaries with total or partial recall completion codes or without a completion code are excluded

TABLE 3 Mean Expenditures by Day of Reporting Period - 1972 Diary

	Day of Reporting Period													
EC	1	2	3	4	5	6	7	8	9	10	11	12	13	14
01	.15	.11	.11	.11	.10	.08	.08	.12	.10	.10	.12	.10	.08	.08
02	.43	.33	.35	.36	.33	.27	.28	.36	.32	.30	.34	.31	.24	.27
03	.72	.65	.67	.51	.58	.43	.42	.68	.55	.53	.57	.50	.38	.44
04	.47	.30	.33	.29	.32	.27	.25	.34	.34	.30	.32	.29	.22	.25
05	.23	.16	.17	.16	.17	.14	.13	.17	.16	.16	.17	.19	.12	.13
06	.24	.15	.15	.14	.16	.12	.11	.19	.15	.13	.16	.13	.09	.12
07	.13	.12	.11	.09	.10	.09	.08	.12	.08	.09	.12	.11	.07	.08
08	.13	.09	.09	.08	.08	.06	.06	.10	.07	.08	.08	.08	.06	.06
09	.42	.28	.29	.29	.27	.24	.25	.33	.27	.27	.29	.26	.23	.25
10	.30	.22	.22	.23	.21	.16	.16	.23	.20	.20	.23	.19	.16	.17
11	.22	.17	.14	.13	.15	.10	.10	.15	.14	.12	.14	.12	.09	.10
12	.24	.16	.17	.16	.15	.13	.12	.17	.16	.14	.16	.15	.11	.12
13	.16	.12	.13	.11	.13	.09	.10	.13	.11	.11	.12	.11	.08	.09
14	.17	.14	.13	.12	.13	.10	.10	.14	.13	.11	.14	.12	.08	.10
15	.16	.12	.12	.12	.12	.09	.09	.12	.11	.11	.12	.11	.09	.10
16	.13	.10	.11	.10	.09	.08	.08	.10	.09	.09	.11	.09	.07	.08
17	.43	.29	.27	.29	.26	.22	.21	.29	.27	.24	.30	.26	.19	.22
18	.41	.36	.31	.32	.31	.26	.25	.34	.34	.28	.34	.29	.23	.25
19	1.49	1.40	1.34	1.39	1.33	1.18	1.12	1.39	1.25	1.31	1.32	1.25	1.19	1.23
20	.42	.33	.36	.39	.36	.26	.24	.33	.34	.30	.35	.31	.27	.27
27	1.86	.80	.73	.80	.61	.56	.76	.93	.64	.72	.71	.61	.63	1.02
33	.64	.45	.46	.45	.38	.38	.37	.49	.45	.40	.40	.41	.32	.37
47	1.63	1.19	1.02	1.06	1.02	.86	.86	1.23	.98	.98	.95	.96	.87	.97
55	.29	.16	.21	.20	.21	.14	.13	.23	.22	.18	.18	.22	.15	.13
63	.49	.35	.36	.35	.30	.26	.25	.39	.33	.30	.32	.31	.28	.28
64	.37	.29	.24	.28	.25	.21	.20	.27	.24	.23	.24	.24	.20	.24
65	.37	.27	.23	.22	.18	.16	.17	.22	.23	.22	.24	.18	.14	.19

Table 4 Week-to-Week Correlation, 1972 Diary, EC Level

EC	ρ	EC	ρ
01	.5348	16	.2431
02	.6428	17	.4155
03	.2365	18	.4856
04	.2450	19	.5431
05	.4867	20	.6611
06	.4268	25	-.1233
07	.3130	26	.0398
08	.4039	27	.1390
09	.5602	33	.3283
10	.5888	47	.2988
11	.2660	55	.1259
12	.5056	63	.5715
13	.3012	64	.1446
14	.3337	65	.4300
15	.2294		

THE 1972-73 U. S. CONSUMER EXPENDITURE SURVEY: A PRELIMINARY EVALUATION

(Robert B. Pearl, Survey Research Laboratory, University of Illinois)

A large-scale national survey of consumer expenditures was conducted in 1972-73 by the Bureau of the Census on behalf of the Bureau of Labor Statistics (BLS), primarily for the purpose of updating the weights and the selection of items for the Consumer Price Index. An entirely different methodology was used in the 1972-73 survey from that employed in previous BLS undertakings in this field. In the prior surveys, the most recent in 1960-61, the procedure followed was the so-called "annual recall" method. In extremely lengthy interviews, lasting up to 8 to 12 hours although obviously completed in more than one visit, an effort was made to determine the expenditures of the family, large and small, for the entire preceding calendar year. A modified procedure was followed to obtain details about food expenditures and a few other categories by inquiring about such outlays in the week preceding the interview.

The new approach used in the 1972-73 survey attempted to take account of the experience in other survey undertakings aimed at controlling response errors. A number of the techniques were borrowed from the methodologies in use in expenditure surveys in other countries and in university and market research in the U. S.

The survey comprised two major components:

(1) An interview panel consisting of about 10,000 households each year which was visited on a quarterly basis primarily to obtain the larger items of expenditure and certain repetitive items (rent, utilities, etc.) Particular categories were covered either quarterly or on a semi-annual or annual basis, depending primarily on expenditure size.

(2) A diary operation consisting of about 200-250 households per week asked to keep a diary or record of all expenditures for the subsequent two-week period. Although the main focus of the diary was the smaller items of expenditure, the fact that all categories were covered provided various options in compiling estimates as well as many research opportunities.

Because of the critical uses which will be made of the data and the marked change in methodology, an evaluation of the results is clearly in order. This may be especially important in view of current plans to institute a continuing survey of consumer expenditures using a similar methodology, in place of the intermittent efforts which have characterized this field in the past. The purpose of this paper is to present the preliminary results of such an evaluation, which should also be relevant for other survey endeavors which use or could use similar techniques. 1/

1/ A detailed report on the evaluation will be issued in the Census Bureau's Working Paper series. The research has been performed under a joint statistical agreement between the University of Illinois and the Census Research Center for Measurement Methods.

Description of survey procedures

Some additional details about the survey procedures may be useful in following the discussion in the remainder of this report. In the diary operation, a given record book contained space for recording expenditures for 7 consecutive days A diary was placed by means of a personal visit by an interviewer who returned 7 days later to pick up the first book and leave a second, which was collected the following week. Each diary contained a set of two facing pages for each day of the 7-day period. The left-hand page was devoted entirely to food and beverage purchases for home use and was subdivided into several sections with general product headings (dairy and bakery products; meat, fish, and poultry; fruits and vegetables, etc.). A section was provided at the top of the right-hand page for recording expenditures for meals and snacks purchased in restaurants and other eating places. The remainder of the right-hand page was divided into small sections for various non-food categories, with principal emphasis to the kinds of small, every-day expenditures for which the diary procedure was primarily intended. Obviously, not all products and services could be specifically mentioned, so that a good many were relegated to a catch-all section. Since only one record book was provided for a given 7-day period, it is likely that a single respondent (usually the homemaker) kept the diary for the entire family.

The quarterly panel questionnaire was a document of imposing, if not overwhelming size, although not all items, fortunately, were asked each quarter. The information was collected by personal interview, usually with a single household respondent. The subjects covered each quarter included home repairs and alterations, utility and fuel costs, clothing and household linens, equipment repairs, vehicle repair and maintenance, and trips and vacations, among others. The questioning was conducted on a semi-annual basis for small household appliances and equipment, furniture and other home furnishings, health expenditures, education, and a few miscellaneous items such as catered affairs, funerals, and moving expenses. The topics covered on an annual basis included rent, mortgage payments and other housing costs, major appliances, vehicle purchases, insurance premiums, subscriptions and memberships, and a few others.

A number of special techniques were employed for various expenditure categories in the quarterly panel. One which is relevant to the discussion in this report is the so-called "inventory" approach used for household appliances and vehicles. Instead of inquiring directly about expenditures for a given period, respondents were asked at the first interview about possession of the articles in question. If any such items were present, the date of acquisition was determined and, if within the previous year, the cost and a variety of other characteristics were recorded. The items in the inventory were

differentiated between those purchased by the family for its own use and those it had received as gifts from persons outside the household. This inventory was updated at specified subsequent visits with inquiries about any new acquisitions, differentiated into the same two classes. Questions were also asked at these updatings about items purchased by the family as gifts to be given to persons outside the household. Thus, there were two measures of expenditures for gifts (the reported value of those received and the reported cost of those given), either of which could be used as part of the total expenditure estimate.

Aside from the manner of estimating gifts, another feature of the inventory approach is the possibility of deriving two separate expenditure estimates for each year. The first, which is the one ordinarily used, is a direct estimate of acquisitions in that year derived by updating the inventory in the course of the survey for that year. The second is an indirect measure obtainable from the initial inventory in the survey for the following year, whereby items secured during the previous year can be identified from the reported date of acquisition.

Summary of findings for expenditure categories

In the remainder of this paper, an assessment is attempted of the adequacy of the expenditures data obtained in the 1972-73 survey for the various categories of goods and services. The general approach used in this appraisal has been to compare the estimates from the quarterly panel with those from the diary operation, where the same subject was covered in both, and to relate either or both to various independent sources of expenditure data. The principal objectives are to assess which of the survey procedures appeared to be more effective for particular categories of expenditures and to determine what types of improvements and modifications may be suggested by the results. The conclusions can only be tentative because of major uncertainties about the validity and comparability of the independent data used as a standard and because adequate detail was often unavailable to explore the subject in sufficient depth. Nevertheless, in a substantial number of cases, persistent patterns emerged across category lines which pointed in rather specific directions.

The most frequently used of the independent data sources are the Personal Consumption Expenditure (PCE) estimates prepared by the Department of Commerce in conjunction with the Gross National Product Accounts. These represent, essentially, the market value of goods and services purchased by persons and nonprofit institutions in the U. S. The estimates are developed from a variety of primary data sources by means of a o complex series of transformations, the reliability of which is indeterminate. Moreover, the PCE data are compiled only in summary form on a current basis. Detailed estimates are provided only for benchmark (quinquennial economic census) years with the most recent available at this writing relating to 1967. In order to derive the necessary level of disaggregation for these comparisons, it was necessary for the author to update the detailed PCE estimates from 1967 to 1972 using appropriate Census of Manufactures and Census foreign trade data, a step which of course adds to the uncertainty. The other independent data derive mainly from Government administrative, census, or survey sources although some private sources are also used.

It should be noted that the survey results used in this evaluation are derived from special tabulations of re-weighted original data tapes. They do not reflect editing changes which may have been made at later stages of processing by BLS. As a result, the figures may differ somewhat from those already published or to be published by that Agency or which may be compiled from the public-use data tapes recently issued. Certain differences in time reference and conceptual approach would also contribute to the disparities.

Table 1 presents a summary of the findings for the various expenditure categories. For purposes of summarization, a number of the detailed categories have been combined and averaged. The table designates the "best" survey source, that is, the one generally closest to the independent data, in cases where the two survey estimates are significantly different. The ratios of the "best" survey estimates to the independent estimates are indicated in terms of broad class intervals, allowing insofar as possible for conceptual differences between the sources, but the actual computed values are also provided.

1. Food and beverage expenditures--After allowance insofar as possible for various conceptual incomparabilities, there appeared to be a reasonably close correspondence between the diary estimates of food purchases for home use and the independent sources. The fact that the homemaker --the usual diary keeper for the family--is ordinarily responsible for most of the purchases was undoubtedly a positive factor. The allocation of maximum space on the diary record to this expenditure class probably contributed as well.

There were considerable disparities, however, in the precision with which various food categories were reported. The reporting was apparently most complete for relatively costly items, such as meat and poultry, and for those used promptly and on a daily basis, such as milk and other dairy products and bread and fresh-baked items. The coverage seemed to be considerably less complete for food staples such as flour, shortening, and sugar which are bought less frequently, with each purchase used over a considerable period of time. One of various possible explanations for these differences is that many respondents may not start keeping their diaries promptly--or do not make entries, as requested, on a daily basis--but later attempt to reconstruct the omitted periods by memory. In doing so, items which represent the main course in a meal or which are purchased and used relatively frequently might be more readily recalled.

A less anticipated finding was the close correspondence between the survey and independent estimates for meals in restaurants or other eating places, where a substantial proportion of

the outlays would be made by individual family members other than the homemaker. The prominent positioning of the section for reporting purchased meals on the diary record and some emphasis to this subject at the time of the diary checking procedure might have contributed to this outcome. At the same time, the marked deficiency for alcoholic beverages confirms the continued failure of household surveys to measure a sector where there is considerable sensitivity about reporting.

2. Small expenditures other than food--For various small expenditure items other than food, for which the diary was the principal if not only source, a predominant factor appeared to be the role of the various family members in making purchases. Where the responsibility was principally that of the homemaker, such as for laundry or cleaning products or household services, the reporting appeared to be considerably more complete than in cases where other members were substantially involved, as for toiletries or hair care. Even for those expenditures where the homemaker predominated, however, the reporting appeared to be generally less adequate than for food purchases, probably partly a reflection of the much smaller amount of space and attention given to non-food items on the diary record.

3. Clothing expenditures--As was anticipated to some extent, this expenditure category represented one of more troublesome sectors, with neither survey source exhibiting any clear cut overall advantage and neither corresponding very closely with the independent data. Following the pattern observed throughout the analysis--and expected from previous experience--the larger items (suits, coats, etc.) were apparently more adequately reported and the quarterly panel emerged as the superior source in this case. Also not surprisingly, the diary procedure represented the "best" source for a diversified category such as accessories, where it was probably difficult to communicate the full range of items in an interview procedure. For no apparent reason, the diary estimates also provided the closer correspondence with the independent data for footwear, although this subject was probed in much greater detail in the quarterly panel.

For the broad range of middle and lower priced clothing products, the advantage seemed to alternate between the two survey sources, without any consistent relationship to the importance of the item. One problem which complicated the appraisal--and which extended to most other expenditure classes as well--was the existence of a large residual clothing group in the diary estimates, consisting mainly of incomplete or inadequate entries which could not be assigned to specific categories.

4. Household appliances--The results for both major and minor household appliances, for which the quarterly panel was the rather evident source, represented one of the more successful outcomes of the survey. The "inventory" approach, described earlier, was evidently an important factor in this showing and it appeared that the benefits could be maximized when the technique was used to its fullest extent. For categories where gifts are significant (e.g., small kitchen appliances and sound equipment), this appeared to suggest using, for purposes of estimation, the estimated value of gifts received by the family from others (which derived directly from the inventory of items on hand) as opposed to the reported cost of gifts given by the family to persons outside the household, which was based on a recollection of previous purchases. Also, there was some evidence in favor of pooling the two estimates obtainable for a given year under the inventory approach, the one based directly on the survey for that year and the other derived indirectly from the initial inventory in the survey for the following year. Comparisons between those two sets of estimates indicated no evident superiority of either over the other in relation to the independent sources. Pooling of the estimates would have roughly the effect of doubling the sample size, a considerable advantage for items with especially large variances.

5. Household furnishings--This broad category provided a rather clear-cut demonstration of the relationship between the size of an expenditure and the likelihood of its being reported. The closest correspondence with the independent estimates was found for furniture, the most costly class, followed by the next most significant group--floor, window, and furniture coverings. In both cases, the quarterly panel appeared to be the superior source, partly on the basis of sampling variances. The survey estimates fell considerably short for household linens and especially for smaller products such as dinnerware and cookware, luggage, and decorative items. In these latter instances, the diary estimates, although themselves deficient, appeared to be at least equivalent in coverage to those from the interview panel.

6. Automobile and vehicle expenses--In nearly all cases, except for the diversified accessories group, the survey data corresponded rather closely with the independent estimates. Moreover, although the homemaker would normally have less responsibility for this sector than most others, the diary estimates were not significantly different from those from the interview panel for some of the smaller categories, particularly gasoline and accessories. The quarterly panel was, as usual, clearly more effective for the larger items--vehicle purchase, tires, and insurance--and also appeared to provide somewhat more complete results for maintenance and repair expenditures.

7. Housing expenditures--Once again, with the exception of one rather diffuse category (fuel purchases), the survey and independent estimates corresponded rather closely. Also, in the main, the diary-based data, again unexpectedly, matched those from the interview panel.

8. Health expenditures--One of the more pleasant surprises was the relatively close correspondence between the survey and independent estimates for most health expenditures, usually considered to be one of the more treacherous

areas in this kind of undertaking. A less optimistic reading might be that the survey results were at least as good as most previous endeavors of a similar nature, without attempting to categorize their accuracy in an absolute sense. In any event, the diary procedure once more provided the most unexpected outcome, in that not only for small items such as drugs and medicines, but also for most professional health services, the data seemed to hold up surprisingly well. An exception was for hospital services but even the quarterly panel data in that instance seemed somewhat deficient, possibly because of complications introduced by the pervasive role of health insurance and other third-party payors.

Conclusions and Recommendations for Diary Procedure

Many of the survey findings were predictable on the basis of previous experience in this field, but a rather unexpected outcome was the relatively effective performance of the diary operation in some sectors, such as housing and health, where the quarterly panel had been assumed to be the only realistic option. At the same time, a number of deficiencies were observed in the diary results even in categories where that procedure was deemed to be the more appropriate source. In addition to the specific comments made earlier for particular expenditure categories, the general conclusions affecting the operation are as follows:

1. Use diary procedure as primary source unless evidence clearly dictates otherwise--A general rule of thumb suggested by the findings is that, unless a clear-cut reason exists for using an interview procedure such as the quarterly panel, dependence might better be placed on the diary approach for a given expenditure category. The diary would be a dubious source for items with exceptionally large variances (vehicles, appliances, furniture, etc.) or where unusual payment arrangements might require special questioning (insurance paid through payroll deductions, mortgage payments made automatically through bank accounts, hospital bills paid largely but belatedly by insurance, etc.). In most other instances, the diary procedure appears to be at least as good a risk as the interview approach, and probably a less costly one as well. A number of modifications and improvements in the diary procedure are clearly necessary, however, in order to overcome some rather evident deficiencies.

2. Limit the range of items any one family would be asked to report--The use of a diary covering all items of expenditure, as was done in the 1972-73 survey, may have certain theoretical benefits, but considerably proscribes the ability to improve the overall process sufficiently to satisfy the expanded requirements just cited. Evidently, as stated earlier, one of the reasons for the more successful coverage of food purchases than other small items in the 1972-73 diary was the much greater amount of space and attention accorded the former. Moreover, for all categories including food, there was a considerable undifferentiated residual group, resulting mainly from incomplete or inadequate entries which could not be classified in detail, which detracted materially from the usefulness of the results. The general lack of space and the inability of interviewers to focus on so wide a range of items in reviewing the diaries probably largely accounted for this latter deficiency.

Since it would be impracticable to provide adequate space and annotation for all categories on a single form, one rather evident solution would be to limit the range of items which any one family would be asked to report. For example, one subsample might be asked to report only on food and other supermarket products, a second on clothing and household linens, a third on health-related expenditures, etc. There is obviously some practical limit to the number of subsamples that could be simultaneously operated and a good deal of thought and some experimentation would be necessary to devise a workable plan.

Probably even more important than space considerations, the use of this kind of specialized approach would make it feasible to provide for a more focused set of check questions and procedures at the time of diary pickup to overcome some of the disparities noted in the present survey (such as underreporting of certain food items relative to others). In fact, a modified procedure would likely entail much more of a combination of interviewing and record keeping than is now the case.

3. Vary length of record-keeping periods--As previously noted, the 1972-73 survey provided for two weeks of record keeping for each sample family, covering all items of expenditure. If specialized subsamples are developed as proposed above, it is obvious that either a larger overall sample would be needed or much higher sampling variances would have to be accepted. One way out of this dilemma would be to vary the length of the record-keeping period depending on the variances of the subjects covered for a given subsample. For example, for a low variance category such as food, it might even be possible to reduce the record-keeping period to one week or 10 days. For most categories, however, such as clothing expenditures or health costs, an increase in the period of record keeping to up to 3 months or longer might be considered. The fact that only a limited set of items is covered might reduce the reporting burden sufficiently to secure extended cooperation of this kind. In such a system, use of less costly collection methods such as having respondents mail in completed diaries on a periodic basis (monthly, semi-monthly, etc.) would be more practicable.

4. Provide separate diaries, where indicated, for individual members--As noted, only one person, usually the homemaker, probably maintained the diary for the entire family in the 1972-73 survey. Not surprisingly, the results were clearly more favorable for the kinds of expenditures for which the homemaker was mainly responsible than for those likely to be made by other members. One possible way of obtaining more consistent results, where the expenditures to be

reported are of a more dispersed nature, would be to provide separate diaries for all family members above a certain age (perhaps 12 and over) on which to record their individual disbursements. For this purpose, the diaries could be briefer and less formal than the main record for the family.

5. Reconsider matter of providing monetary or other incentives for cooperation--An experiment was conducted in the early stages of the 1972-73 survey on whether an offer of cash payments would materially improve cooperation in maintaining diaries. The results were inconclusive in this regard and the incentives were dropped from the procedure. However, most previous experience supports the notion that both cooperation and adequacy of reporting are benefited by some inducements of this kind. If greater dependence is to be placed on diaries in a continuing operation, as is being proposed, it would seem especially important to reconsider this matter of offering incentives and to experiment with alternative approaches.

6. Continue exploration of timing biases--Although not mentioned up to this point, perhaps the most conclusive survey finding was affirmation of the traditional bias found in diary operations, whereby a higher level of expenditures is reported in the earlier as opposed to the latter stages of the record keeping period. In a 2-week diary procedure, for example, the estimates for the first week are almost invariably higher than those for the second week. Differences of this nature were found for virtually every expenditure category in the 1972-73 survey. The margins of difference varied a good deal, however, and seemed to be almost random in nature. There was no evident relationship of the weekly differences for a category to other survey measures, such as the completeness of reporting relative to the independent estimates.

Many explanations have been offered for this phenomenon. One theory suggests that expenditures are exaggerated in the early stages because of "telescoping," that is, inclusion of some purchases made prior to the record-keeping period. Another is that reporting tapers off toward the end because of fatigue. Temporary alteration of buying habits because a diary is being kept has even been suggested. More detailed information on this subject, which will be presented later, will hopefully shed further light on this matter, although it is unlikely that the issue will be fully resolved. It is possible that some changes in survey procedures will be suggested by these findings, but the scope and nature of these are still uncertain.

7. Explore use of universal product codes--A promising technological development that could affect future survey work in this field is the inclusion of "universal product codes" on most canned and packaged supermarket and drug store items (and likely to extend to many others). A useful experiment would be to ask respondents to record these codes in their diaries where available, as well as brief product descriptions, to assess how accurately this information is reported. If the effort is sufficiently productive, respondents might be relieved of the necessity of describing products in any detail where code numbers are available. More importantly, this step could result in far more accurate and consistent classification of products reported in surveys and a major reduction in the coding effort required at the data processing stage.

Table 1 -- Summary of Findings for Expenditure Categories: 1972-73 Expenditure Survey Data Compared to Independent Sources

(Table condensed because of space limitations; full detail in forthcoming report, see text footnote 1)

Category	"Best" Survey Source 1/ QP= Quarterly Panel D= Diary Operation N= No Significant Difference	Ratio of "Best" Survey Estimate to Independent Estimates: Best Judgment as to Range of Ratios (allowing for conceptual and other differences between sources) 2/ A= .9 or higher B= .8 to .89 C= .7 to .79 D= .6 to .69 E= less than .6	Survey Results: Actual Ratio of Survey to Independent Estimates	Survey Results: Estimated Standard Error of Ratios	Principal Independent Sources Used (PCE= Personal Consumption Expenditures, GNP Accounts) 3/
Food Purchases for Home Use*	D**	A,B	.86	< .01	PCE, Dept.
Meat, Milk, Bread, etc.		A	.96	.02	of Agricul-
Food Staples (Flour, Sugar, Oil)		D	.65	.02	ture
Purchased Meals and Snacks*	D	A	1.10	.02	PCE, Census
Alcoholic Beverages*	D	E	.38	.03	of Business
Small Non-Food Expenditures	D**				PCE, Census
Mainly Responsibility of Homemaker					of
Goods (Laundry-Cleaning Products)		C	.72	.02	Selected
Services (Laundering, Domestic Help)		A	1.04	.02	Service
Dispersed Responsibility					Industries
Goods (Toiletries, Film, etc.)		D,E	.57	.02	
Services (Hair Care, Sports, etc.)		C	.68	.02	
Clothing Expenditures*	N	C	.73	.01	PCE, Market
Larger Items (Coats, Suits)	QP	A	1.11	.03	Res. Corp.
Medium and Smaller Articles	N	C	.76	.01	of America
Accessories (Hats, Ties, etc.)	D	C	.72	.03	(MRCA)
Footwear*	D	B,C	.79	.01	Cens. of Bus.
Household Appliances	QP**				PCE
Major Appliances		A	1.00	.02	
Minor Appliances		A	1.01	.03	
Home Furnishings					
Furniture*	QP	A,B	.91	.03	PCE, Census
Floor Coverings, Drapes, etc.	QP	B,C	.75	.03	of Bus.,
Dinnerware, Luggage, Tools, etc.	N	E	.48	.02	MRCA
Vehicle Expenses					
Vehicle Purchase*	QP**	A	1.01	.03	PCE, Annual
Gasoline and Oil*	N	A	.98	.01	Housing
Accessories, exc. Tires*	N	D	.67	.02	Survey,
Repairs and Maintenance*	QP	A,B	.88	.02	Cens. of Bus.
Housing Expenditures					PCE, Annual
Rent*	N	A	1.03	.02	Housing Surv.,
Mortgage Payments, Taxes*	QP	A	1.02	.04	Survey of
Home Repairs and Alterations	N	A	1.04	.06	Alterations
Utility Bills	N	A	1.02	.02	and Repairs
Fuel Costs (Fuel Oil, Coal)	N	C,D	.61	.02	
Health Expenditures					Soc. Security
Hospital Services*	QP	C	.76	.04	Adm, Center for
Professional Services*	N	A	.98	.02	Health Adm.
Drugs and Sundries	D	A	1.02	.05	Studies, PCE

* Signifies comparisons based on two years of data; other comparisons based on 1972 only.

1/ The "Best" survey estimate is defined as the one--quarterly panel or diary--closest to the independent figures. Where a double asterisk (**) is appended to the code, this indicates the specified source was the only one for which comparisons with the independent data could be made in the required detail.

2/ The ranges in this column are not always entirely consistent with the computed ratios in the next column, but make allowances for conceptual differences between survey and independent sources, disparities among the independent sources, and other factors for which numerical adjustments cannot be made.

3/ Detailed description of sources in forthcoming report, see text footnote 1.

INCENTIVE EFFECTS ON AMOUNTS REPORTED IN AN EXPENDITURE DIARY SURVEY

Charles D. Cowan, Bureau of the Census

Review

Various attempts have been made to imbue survey respondents with an appreciation of the importance of their individual contributions to the response rate by offering gifts or incentives. These attempts are measures designed to increase response to surveys or to improve the quality of response. Many attempts have been made to measure the effectiveness of incentives. A novel experiment by Chromy and Horvitz [4] offered set and variable incentives, the variability arising from the degree to which the respondent wanted to get involved; the higher the involvement, the higher the incentive payment. The variable incentive plan was adopted for use in later waves of the study because it was shown to be the most cost effective.

Several studies have analyzed the effects of offering incentives in mail surveys. Generally, the payment of incentives was found to improve results, and prepayment of an incentive more rewarding to the sponsor than promised payment. Linsky [14] in a review of the use of incentives in mail surveys states that "cash rewards invariably increased returns over the level of response for no-reward control groups" in ten experimental studies. His interpretation of these results suggest that the "motivating power of the (incentive) is not in terms of its monetary value but in its symbolic, or token value". In the area of personal interview surveys, the use of incentives has come about either as a curative for or a preventive against unacceptably low response rates. Ferber and Sudman [8] reviewed the effectiveness of compensation in consumer expenditure surveys, finding variable results. They hypothesized that the success of offering compensation could be due to the auspices under which the incentive is offered, the income level of the recipient, or the subject of the study.

In other diary surveys, Sudman[22] found no important differences in recording levels due to the offer of compensation, and no change in reporting levels when the level of compensation is changed in followup contacts. But in a later study, Sudman and Ferber[21] found that the offer of compensation improved the level of cooperation 14 to 17 percent. Kemsley and Nicholson[13] in England also found an increase in response rates as a function of the amount of monetary compensation offered.

For the most part, the use of incentives seems to be effective in raising response rates, especially in mail surveys, but also in diary and panel surveys, where the impetus is to encourage respondents in a task requiring more commitment than the usual one-time interview. The next section of this paper will report on an attempt by the U.S. Bureau of the Census to secure cooperation on a diary survey.

The 1972-73 Survey of Consumer Expenditures

Sudman and Ferber[21] point out in their review that very little work has been done to validate results in compensation experiments. This paper attempts to assess the effectiveness of compensating respondents, and what changes may be measured in item response rates and in the frequency of reporting of expenditures. Work done in the area has generally examined only the improvement in the response rate as a measure of the effectiveness of offering incentives. And in most diary studies validation is extremely difficult, so the efforts undertaken make the assumption that "more is better", since the major problem in diary surveys is underreporting. Since validation is very difficult, this paper will also make the assumption that more is better, trusting in the efforts of earlier researchers in this field. An earlier report (Walsh[23]) was made of the response rate differences for this survey due to use of incentives. The results show no significant differences in the overall response rate to the survey, the rate climbing from 72 percent to 77 percent, the standard error of the difference being approximately 5 percent.

The Consumer Expenditure Survey (Pearl[15]) was conducted by the Census Bureau for the Bureau of Labor Statistics in calendar 1972 and 1973 to obtain comprehensive information on consumer expenditures to revise the National Consumer Price Index. The diary was a stratified multistage design consisting of 13,500 sample units in each year distributed among 30 self representing primary sampling units. The unit of analysis was the consumer unit (CU), which is essentially a family, or a group of people living together who pool their resources. The sample was systematically divided into 52 weekly subsamples to control for seasonality. In anticipation of difficulties in securing respondent cooperation in completing diaries in each of the two weeks a CU was in sample, an experiment was designed to test the effectiveness of offering an incentive to sample households for participating.

The experiment was conducted for eight weeks of the survey. The treatments were to offer no payment to a randomly designated one third of the sample, five dollars to one third, and ten dollars to the remainder of the sample. Each interviewer handled only one treatment since it was felt that the burden imposed on the interviewer of keeping track of incentive offerings would be too great. The sample consisted of about 1850 eligible units, with 1472 units completing one or both diaries as requested. This experiment was admittedly a very small scale effort, but the results after eight weeks were considered definitive enough to terminate the experiment with the decision not to offer incentives.

Results

Table 1 presents results for 72 expenditure categories covered by the diaries. The figures in the first three columns represent average weekly expenditures in dollar amounts reported by the respondents under the varying levels of payment. Column four shows the F statistic from a one-way analysis of variance (ANOVA) on the mean level of expenditures. An F value of 3.00 would be significant at the 95 percent confidence level, given 2 and ∞ degrees of freedom. The degrees of freedom for the denominator of the test actually varies, since the number of observations varies

due to missing observations for various expenditure categories, but is always greater than 2800. The last column, eta-squared, reports the proportion of variance explained by the differences in the means between payment levels.

Three assumptions underlie the use of the F-statistic in analysis of variance: equal variances between treatments, normality of observations on the dependent variable, and independence of observations. To test for equality of variances, Cochran's test[1] was used. Of the 72 expenditure categories, only 3 showed an indication that the variances were different, and these items were items which were infrequently bought (in any one week period), so that a few households reporting purchases of an item may cause a change in the variance estimated from a relatively small sample.

Regarding normality, the assumption depends on the expenditure category. Categories like total food expenditures and total other expenditures appear to be normally distributed, though no formal test was employed to make the determination. In categories involving infrequent expenditures, like medical expenditures, the distribution is bimodal, with a number of reports of zero expenditures for those respondents with no purchases of the item, and the remainder of the observations being actual expenditure amounts. ANOVA has been shown to be robust under departures from normality, and in view of the results shown later, it is not believed that the departure from normality is serious.

The final assumption is independence of observations. The sample consists of 1472 consumer units drawn at random in a two-stage stratified sample, each unit filling out a diary in each of two successive weeks. This gives 2944 observations total, except that many respondents failed to fill in various sections of the diary, so the number of observations ranges between 2835 and 2944. Sample units are independent, and reports within the sample unit for the first and second week of reporting evidence a low level of correlation. Observations are treated as independent for the purpose of this analysis[2].

The statistic eta-squared is computed as the ratio of the explained sum of squares to the total sum of squares. Eta-squared is used in this paper only for descriptive purposes, to make the point that statistical significance, estimated by the probability of a Type I error less than 5 percent, does not necessarily indicate meaningful or substantial results that would lead to the adoption of the use of incentives. In no case is even one percent of the variance explained by the payments.

Table 2 presents the proportion of zero responses to an expenditure category. A zero response means that the respondent did not report buying an item in the diary. To further the analysis of people's response to incentives, a device will be employed here to focus on a specific type of behavior. Changes in reporting observed between treatments can be due either to a respondent reporting a purchase where he otherwise would have reported no purchase, or a respondent who was reporting purchases, now reporting more purchases under the influence of the incentive payment. However, the changes in the mean expenditures between treatments can be examined as a function of the decrease in the proportion of zero responses to a question, and the change between means for those who do report purchases. The former case can be taken as a nonresponse or the tacit reporting of no purchase. Were the proportion of zero expenditures to decline as a function of the incentive payment, then one could conjecture that some underreporting due to a lack of effort by the respondent was lessened by use of incentive payments.

A careful study of table 2 shows that for most items the proportion of zero responses does indeed go down, at least between the nonpayment and the combined five and ten dollar treatment groups. For the first category, Total Food and Beverages at Home, a comparison[3] of the three proportions of nonreporters yields a Chi-Square of 14.02 ($\alpha < .001$) with two d.f., and a z-test[4] between the respondents receiving no payment and respondents receiving either five or ten dollars is 3.38 ($\alpha < .001$), so it appears that incentives may increase the number of reports in the diary.

An examination of levels of expenditures, again for Total Food and Beverages at Home, with the zero expenditure reports removed yields average expenditures of $21.40, $22.69, and $23.15 for no payment, five, and ten dollars compensation respectively. The zeros removed may have been legitimate nonpurchases, but what is being analyzed here is an increased reporting of expenditures. The range between low and high expenditure levels between treatments has been reduced. The F value for the difference in the above means is 2.05, not significant at $\alpha = .05$, and the proportion of variance explained by the different levels of compensation was half of the variance explained by the original model in Table 1 (.0017 of .0036). This analysis was not extended to other categories because of the relatively low response rates to individual items. Besides the low purchase rates, only 14 of the 72 categories showed statistically significant improvement, and some of these tests were correlated with one another because some categories are aggregates of several others. The other direction of interest in the analysis is whether response to incentives is interactive with any variable that may be used in stratification of the sample. A number of demographic variables were used in the analysis presented in Table 3, where Total Food and Beverages at Home is the dependent variable in a two-way analysis of variance, with incentives and one of the demographic variables as the independent variables in each analysis. The first three columns of Table 3 are mean expenditures for Total Food and Beverages at Home for the different incentive levels adjusted[5] for the other variable in the analysis. The adjustment of the means is calculated as the deviation from the grand mean estimated for the row variable after the effects of the column variable have been removed. The adjustment process accounts for any correlation between incentives and the demographic variables. For example, the first line in Table 3 presents the mean expenditures unadjusted. The second line presents the means adjusted for urbanicity.[6] The next four columns are F values for the main effects, the incentive treatments,

the other variable in the two-way analysis of variance listed at the left hand side of the table, and the interaction between the incentive variable and the demographic variable on the dependent variable. The main effects represent the linear effect of incentives and the demographic variable in the analysis. Because of the missing data and the vagaries resultant from lack of control in sampling housing units, incentives is slightly correlated with each of the demographic variables, and so main effects accounts for the predictive power of the two independent variables. The final two columns give the eta-squared values for the incentives and the dependent variable. All F-values for main effects, incentive (row) effects, and column effects are significant ($\alpha < .01$), and F-values for the first six interaction effects are significant ($\alpha < .01$ except for the age-sex combination, $\alpha < .02$). The next section will consider if these effects are meaningful.

Conclusions

In the above presentation, F-values and their significance levels have been dutifully presented but without commentary regarding the interpretation of the results. The F-values generated are"significant" but not too exciting in the sense that none of the items examined displayed an overwhelming response to the payment of incentives. The small differences between incentives groups noted in Table 1 were found to be an increased reporting on the part of those respondents already listing expenditures without incentives. Additionally, there does seem to be some interaction between incentives and certain demographic variables, meaning some subgroups of respondents are more responsive to incentives than others. But if the reader will refer back to Tables 1 and 3, the eta-squared values show much less than one percent of the variance of reporting explained by incentives, and no more than 1.2 percent of the variance explained by the interaction of incentives and the demographic variables. These minor improvements signified by the eta-squared values seem to indicate that it was not worth the cost of paying incentives to the respondent in this survey, especially when one considers that the overall response rate to this survey did not change significantly.

The expenditure to be made for incentives payments in the full-scale survey were undoubtedly better spent on other response improvement techniques outlined by Walsh [23]. Better training of interviewers to improve respondent commitment to completing the diary provided results as good as, and hopefully better than, those obtained above. A system of telephone or postcard reminders in the middle of the diary week might also improve response in future efforts.

Because of the general favorable results other researchers have found, future experimentation with use of incentives may yield more satisfying results. One possible experiment would be to administer the diary to all sample households, and offer an incentive to those refusals in a random half-sample to determine if response rates and reporting behavior differ significantly.

Table 1: Average Weekly Expenditure Levels by Amount of Incentive

Expenditure Category	Payment $0	$5	$10	F[1]	Eta[2]
Total Food at Home 2/	$17.84	$20.21	$20.16	5.2	.0036
Cereal & Cereal Prod.	.47	.57	.59	4.8	.0034
Bakery Products	1.74	1.92	1.94	2.4	.0017
Meat	5.40	5.71	5.86	.9	.0006
Poultry	.72	.81	.74	.8	.0006
Fish & Seafood	.40	.71	.52	5.4	.0037
Comb. Meat & Poultry	.00	.00	.00	.2	.0001
Dairy Products	2.67	2.87	2.94	2.3	.0016
Milk, Cream & Milk Prod.	2.62	2.82	2.85	2.0	.0014
Other Dairy Products	.05	.05	.08	2.8	.0019
Fruits	1.55	1.61	1.54	.4	.0003
Fresh Fruits	.95	1.03	.97	.8	.0005
Frozen Fruits	.02	.01	.01	.6	.0004
Canned & Dried Fruits	.23	.20	.21	.6	.0004
Fruit Juices	.35	.36	.34	.2	.0001
Vegetables	1.49	1.54	1.65	1.6	.0011
Fresh Vegetables	.99	1.01	1.05	.5	.0003
Frozen Vegetables	.14	.14	.17	1.4	.0010
Canned & Other Vegetables	.38	.39	.42	.7	.0005
Sugar & Other Sweets	.52	.58	.61	2.3	.0016
Nonalcoholic Beverages	1.63	1.76	1.91	4.1	.0028
Carbonated Drinks	.88	.94	1.06	3.4	.0023
Other Nonalcoholic Bev.	.76	.82	.84	1.2	.0008
Baby, Junior & Toddler Food	.04	.17	.10	7.6	.0051
All Other Food at Home	2.76	2.84	3.04	.8	.0006

Table 2: Proportion of Respondents Who Reported No Expenditure For A Category by Amount of Incentive

	$0	$5	$10	Total
Total Food at Home	.166	.109	.129	.134
Cereal & Cereal Prod.	.580	.516	.488	.528
Bakery Products	.213	.177	.184	.192
Meat	.246	.229	.211	.229
Poultry	.688	.680	.661	.677
Fish & Seafood	.737	.676	.683	.699
Comb. Meat & Poultry	.998	.997	.998	.998
Dairy Products	.176	.147	.156	.159
Milk, Cream & Milk Prod.	.178	.154	.159	.163
Other Dairy Products	.944	.941	.925	.937
Fruits	.330	.321	.317	.322
Fresh Fruits	.429	.443	.431	.435
Frozen Fruits	.986	.983	.983	.984
Canned & Dried Fruits	.769	.763	.764	.765
Fruit Juices	.674	.654	.686	.671
Vegetables	.327	.312	.287	.309
Fresh Vegetables	.399	.384	.368	.384
Frozen Vegetables	.827	.837	.779	.815
Canned & Other Vegetables	.634	.631	.597	.621
Sugar & Other Sweets	.539	.500	.450	.498
Nonalcoholic Beverages	.325	.285	.265	.291
Carbonated Drinks	.511	.491	.439	.482
Other Nonalcoholic Bev.	.533	.519	.479	.511
Baby, Junior & Toddler Food	.965	.945	.960	.956
All Other Food at Home	.245	.213	.186	.215

Table 1 (Continued): Average Weekly Expenditure Levels by Amount of Incentive

Expenditure Category	Payment $0	$5	$10	F	Eta[2]
Total All Other Expenditures	$134.01	$116.20	$123.42	.9	.0006
Food & Bev. Away from Home	6.48	7.00	8.25	6.5	.0044
Personal Care Products	1.28	1.37	1.65	3.9	.0026
Personal Services	1.23	1.17	1.29	.4	.0003
Household Supplies	2.41	2.62	2.78	2.2	.0015
Housekeeping Services	2.99	3.33	3.50	.4	.0003
Household Help	1.17	1.67	1.73	1.2	.0008
Laundry & Dry Cleaning	.90	.90	.91	.0	.0000
Other Services	.92	.76	.86	.1	.0001
Housing Costs(Rent,Mortgage)	32.03	22.20	20.32	1.9	.0013
Alterations & Repairs	4.25	3.30	7.93	1.7	.0012
Fuels and Utilities	11.07	9.30	9.82	1.4	.0010
Textile Home Furnishings	1.50	1.37	1.52	.1	.0001
Furniture	2.19	1.29	4.85	2.0	.0013
Household Appliances	2.43	1.84	2.93	.4	.0003
Other Household Equipment	2.52	2.13	2.72	.5	.0003
Household Items	.80	.47	.50	1.7	.0012
Outdoor Items	.42	.34	.26	.2	.0002
Hardware, etc.	.12	.19	.07	.6	.0004
Other	1.18	1.13	1.90	1.5	.0010
Insurance, etc.	.98	.64	.66	.4	.0003
Clothing & Related Items	8.94	9.76	10.75	1.4	.0010
Clothing, All Persons	6.72	7.20	8.03	1.1	.0007
Footwear	1.43	1.70	1.69	.7	.0005
Infant & Toddler Wear	.22	.18	.16	.7	.0004
Other Clothing	.57	.66	.87	.3	.0002
Private Transportation	15.00	15.14	17.55	.1	.0001
Vehicle Purchases	5.81	5.40	6.71	.0	.0000
Gasoline, Motor Oil, Etc.	5.19	6.19	5.85	3.2	.0022
Parts & Equipment	.70	.95	1.57	3.1	.0021
Maintenance & Repair	1.97	1.56	2.07	.5	.0003
Other	1.33	1.14	1.34	.3	.0002
Public & Other Trans.	.92	1.31	1.03	.4	.0002
Medical Care	7.96	9.30	6.30	2.1	.0015
Drugs & Medicine	2.24	1.77	1.73	3.2	.0022
Professional Services	4.45	4.56	3.19	.8	.0006
Other Medical Expenses	1.27	3.06	1.37	4.2	.0029
Reading Materials	.97	1.18	1.12	1.4	.0010
Sporting Equip. Toys, etc.	3.00	2.16	2.77	1.8	.0012
Admission Fees	3.16	3.14	2.72	.3	.0002
Miscellaneous Expenses	4.78	3.92	3.51	.8	.0006
Education	3.11	.58	.85	1.3	.0009
Tobacco	1.90	2.16	2.32	3.3	.0022
Alcoholic Beverages	1.83	2.01	2.49	2.9	.0020
At Home	1.45	1.74	1.71	1.1	.0007
Away From Home	.38	.28	.78	5.4	.0036
All Other Expenses	15.80	11.80	7.03	.7	.0004

Table 2 (Continued): Proportion of Respondents Who Reported No Expenditure For a Category by Amount of Incentive

Expenditure Category	$0	$5	$10	Total
Total All Other Expenditures	.127	.087	.082	.099
Food & Bev. Away from Home	.335	.310	.248	.299
Personal Care Products	.551	.496	.478	.508
Personal Services	.756	.770	.743	.757
Household Supplies	.293	.261	.210	.256
Housekeeping Services	.708	.672	.655	.679
Household Help	.920	.903	.888	.904
Laundry & Dry Cleaning	.781	.759	.748	.763
Other Services	.961	.959	.969	.963
Housing Costs(Rent,Mortgage)	.822	.830	.836	.829
Alterations & Repairs	.855	.850	.843	.850
Fuels and Utilities	.636	.627	.615	.627
Textile Home Furnishings	.842	.835	.810	.829
Furniture	.972	.961	.965	.966
Household Appliances	.962	.958	.966	.962
Other Household Equipment	.727	.727	.707	.721
Household Items	.911	.930	.923	.921
Outdoor Items	.958	.965	.964	.962
Hardware, etc.	.971	.979	.974	.975
Other	.816	.796	.779	.797
Insurance, etc.	.985	.988	.993	.988
Clothing & Related Items	.568	.547	.510	.543
Clothing, All Persons	.639	.640	.576	.620
Footwear	.866	.851	.863	.859
Infant & Toddler Wear	.952	.939	.959	.949
Other Clothing	.946	.948	.920	.938
Private Transportation	.372	.341	.325	.346
Vehicle Purchases	.991	.993	.993	.993
Gasoline, Motor Oil, Etc.	.399	.366	.362	.376
Parts & Equipment	.967	.957	.947	.957
Maintenance & Repair	.934	.920	.908	.921
Other	.799	.805	.756	.788
Public & Other Trans.	.910	.870	.866	.882
Medical Care	.554	.548	.559	.554
Drugs & Medicine	.642	.666	.653	.654
Professional Services	.876	.857	.859	.864
Other Medical Expenses	.898	.875	.891	.888
Reading Materials	.495	.482	.469	.483
Sporting Equip. Toys, etc.	.654	.660	.592	.637
Admission Fees	.722	.703	.651	.693
Miscellaneous Expenses	.639	.656	.683	.659
Education	.971	.983	.980	.978
Tobacco	.574	.494	.508	.525
Alcoholic Beverages	.715	.704	.700	.706
At Home	.744	.729	.743	.738
Away From Home	.945	.949	.902	.933
All Other Expenses	.712	.658	.657	.676

1/ Degrees of freedom are 2 and 2900. The latter figure is approximate due to missing data, but differences in significance for such high degrees of freedom are small.

2/ Subcategories may not add to a category total because of treatment of missing data. All values reflect respondent's answers, with no imputed values.

Table 3: Two Way Analyses of Variance of Total Food and Beverages at Home by Incentives and Demographic Characteristics

Variables	Adjusted Expenditure Means $0	$5	$10	F Values 5/ Main 2/ Effects	Incentive	Variable 3/	Interaction 4/	Eta² Incentive	Eta² Interaction
Incentive Alone(3) 1/	17.84	20.20	20.16	---	5.2*	--	--	.0036	--
Incent.& Urbanicity(5)	17.77	20.28	20.15	4.0*	5.7*	3.3*	4.2*	.0039	.0115
Incent.& Ages & No. of Children(4)	18.02	20.44	19.69	79.5*	5.1*	128.6*	3.7*	.0031	.0067
Incent.& No. of HH Members(6)	18.04	20.54	19.55	107.8*	5.9*	148.3*	2.3*	.0032	.0064
Incent.&Race of Head(3)	17.88	20.21	20.11	7.1*	5.0*	8.9*	6.4*	.0034	.0089
Incent.& Age & Sex of Head(12)	17.77	20.62	19.76	28.6*	7.0*	32.8*	1.7**	.0043	.0117
Incent.& Ed. of Head(6)	17.92	20.21	20.07	7.3*	4.8*	8.3*	3.5*	.0033	.0118
Incent.& Work Exper. of Head & Wife(15)	17.78	20.53	19.84	25.7*	6.7*	28.5*	1.4	.0041	.0119
Incent.& Housing Owned/ Rented(4)	17.69	20.54	19.93	30.0*	6.8*	46.4*	.6	.0045	.0012
Incent.& Income of Consumer Unit(7)	17.82	20.46	19.89	29.4*	6.0*	37.3*	1.2	.0039	.0045

* significant at $\alpha < .01$
** significant at $\alpha < .02$

1/ numbers in parentheses are numbers of categories for that variable

2/ see text for explanation of main effects; in brief, "main effects" is the combined effects of the row and column variables, i.e. the linear effects

3/ F-values for the second variable in the two-way analysis of variance, the second variable being defined in the column on the left hand side of this table.

4/ the F-values for the interaction of the incentives variable and the demographic variable listed to the left

5/ Degrees of freedom for the numerator are as follows:
Main effects: d.f. = 2 + (c-1) Variable: d.f. = (c-1)
Incentives: d.f. = 2 Interaction: d.f. = 2(c-1)

where (c-1) is one less than the number in parentheses listed after the variable name on the left hand side of the table. d.f. for the denominator are approximately 2900.

Footnotes

[1] Dixon and Massey [6], p. 310.

[2] See Scheffé [17], Chapter 10 "The Effects of Departures from the Underlying Assumptions".

[3] Fleiss [10], p. 93.

[4] $$z=(p_0-p_c)/(p_0q_0/n_0+p_cq_c/(n_5+n_{10}))^{1/2}$$

where $p_c=(n_5p_5+n_{10}p_{10})/(n_5+n_{10})$

[5] Andrews, Morgan and Sonquist [1].

[6] Urbanicity is the household's location in a central city, a suburb, or a rural area.

[7] Kendall and Stuart [12].

Bibliography

(1) Andrews, Frank, James Morgan and John Sonquist, Multiple Classification Analysis, Ann Arbor, The University of Michigan, 1967.

(2) Armstrong, J. Scott, "Monetary Incentives in Mail Surveys", Public Opinion Quarterly, Volume 39, No. 1, (Spring 1975), pp.111-116.

(3) Cannell, Charles F. and Ramon Henson, "Incentives, Motives, and Response Bias", Annals of Economic and Social Measurement, Vol. 3, No. 2, (April 1974), pp.307-317

(4) Chromy, James R. and D.G. Horvitz, "The Use of Monetary Incentives in National Assessment Household Surveys", American Statistical Association, Proceedings of the Social Statistics Section. 17th Annual Edition (1974), pp.171-179

Bibliography Continued

(5) Cox, Eli P., "A Cost/Benefit View of Prepaid Monetary Incentives in Mail Questionnaires", Public Opinion Quarterly, Vol. 40, No. 1 (Spring 1976), pp. 101-104

(6) Dixon, Wilfred J. and Frank J. Massey, Jr., Introduction to Statistical Analysis, New York, McGraw Hill Book Company, 1969.

(7) Doob, Anthony N., Jonathan L. Freedman and J. Merril Carlsmith, "Effects of Sponsor and Prepayment on Compliance with a Mailed Request", Journal of Applied Psychology, Vol. 57, No. 3 (June 1973), pp. 346-347.

(8) Ferber, Robert and Seymour Sudman, "Effects of Compensation in Consumer Expenditure Studies", Annals of Economic and Social Measurement, Vol. 3, No. 2 (April 1974), pp. 319-331.

(9) Ferber, Robert, The Reliability of Consumer Reports of Financial Assets and Debts, Urbana, University of Illinois, Bureau of Economic and Business Research, June 1966.

(10) Fleiss, Joseph L., Statistical Methods for Rates and Proportions, New York, John Wiley & Sons, 1973.

(11) Huck, Schuyler W. and Edwin M. Gleason, "Using Monetary Inducements to Increase Response Rates from Mailed Surveys", Journal of Applied Psychology, Vol. 59, No. 2 (1974), pp. 222-225.

(12) Kendall, Mayrice and Alan Stuart, The Advanced Theory of Statistics, Vol. 3, New York, Hafner Press, 1976.

(13) Kemsley, W.F.F. and J.L. Nicholson, "Some Experiments in Methods of Conducting Family Expenditure Surveys", Journal of the Royal Statistical Society, Series A, Vol. 123, No. 3 (1960), pp.307-328.

(14) Linsky, Arnold S., "Stimulating Responses to Mailed Questionnaires: A Review", Public Opinion Quarterly, Vol. 39, No. 1, (Spring 1975), pp. 82-101.

(15) Pearl, Robert B., Methodology of Consumer Expenditures Surveys, Working Paper No. 27, Washington, D.C. Bureau of the Census, March 1968.

(16) Pucel, David J., Howard F. Nelson and David N. Wheeler, "Questionnaire Follow-up Returns as a Function of Incentives and Responder Characteristics", Vocational Guidance Quarterly, Vol. 19, No. 3 (March 1971), pp.188-193.

(17) Scheffe, Henry, The Analysis of Variance, John Wiley & Sons, Inc., New York, 1959.

(18) Schewe, Charles D. and Norman G. Cournoyer, "Prepaid vs. Promised Monetary Incentives to Questionnaire Response: Further Evidence", Public Opinion Quarterly, Vol. 40, No. 1 (Spring 1976), pp. 105-107.

(19) Sudman, Seymour and Robert Ferber, "A Comparison of Alternative Procedure for Collecting Consumer Expenditure Data for Frequently Purchased Products", Journal of Marketing Research, Vol. XI (May 1974),pp. 128-135.

(20) Sudman, Seymour, Wallace Wilson and Robert Ferber, The Cost-Effectiveness of Using the Diary as an Instrument for Collecting Health Data in Household Surveys, Urbana, University of Illinois, Survey Research Laboratory, October 1974.

(21) Sudman, Seymour and Robert Ferber, "Experiments in Obtaining Consumer Expenditures by Diary Methods", Journal of the American Statistical Association, Vol. 66, No. 336 (December 1971). pp. 725-735.

(22) Sudman, Seymour, "On the Accuracy of Recording of Consumer Panels: II", Journal of Marketing Research, Vol. 1, No. 3 (August 1964), pp. 69-83.

(23) Walsh, Thomas C., Selected Results from the 1972-73 Diary Survey, Presented at a seminar sponsored by the American Marketing Association, October 7, 1976.

(24) Whitmore, William, "Mail Survey Premiums and Response Bias", Journal of Marketing Research, Vol. 13, No. 1 (February 1976), pp. 46-50.

AN EXPERIMENT IN DATA COLLECTION: USE OF FINANCIAL RECORDS

F. Thomas Juster, The University of Michigan

Introduction

Economists have long been concerned with the consumption and saving behavior of consumers, starting with the simplest absolute income notions and proceeding on through relative income, permanent income, life cycle,wealth and uncertainty hypotheses.[1] Much of the literature attempts to explain consumption rather than consumer expenditures. That is, the mix of saving between additions to financial assets and real assets (durables-housing) is not of concern to many theories, and that mix is of course the essence of most short-term fluctuations in the spending-saving ratio for the household sector.

In general, economists have used two different types of data to examine saving behavior. Traditional explorations continue to rest heavily on attempts to extract good causal specifications from time series data--a pursuit that is perennially enticing but inevitably frustrating. The problems faced by economists in trying to specify appropriate models when working entirely with aggregate time series data are well-known and do not require extensive elaboration: the difficulties are especially relevant where one is interested in the question of which types of income, and with what lag structure, impact on spending-saving behavior. Depending on what else is included in such equations, it is probably true that lag structures of anywhere from one to fifteen years can be "found" without there being any persuasive empirical evidence that one end of this range is much better than the other. Although most economists probably feel that the relevant lag structures are on the short end rather than the long end of this spectrum, their views are more apt to be related to what are seen as plausible modes of behavior than to convincing empirical evidence. To the extent that future income has a variance as well as a mean, and to the extent that large variances mean high discount rates, it does not seem at all reasonable to suppose that distant and uncertain income prospects would carry much weight in current consumption and saving decisions, to say nothing of the fact that imperfections in capital markets sharply reduce the ability of households to borrow against even highly certain (subjectively) future prospects.

Economists have also used cross-section or budget-type data to analyze consumer spending and saving. While only a limited number of such data bases exist, they have been thoroughly mined by the profession. A good deal of the effort devoted to cross-section analysis of consumption data has been given to explaining why the parameters of variables like income look so different in cross-section and time-series analysis. A substantial literature, centering around the permanent income theory, has been built on analysis of measurement errors, transitory changes, and similar phenomena.[2] In general, the budget data have tended to reflect annual consumption and income observations for a cross-section of household. Thus it has been dificult to generate models with any dynamic content, since the basic data lack any good measures of change over time.

It has been increasingly evident that existing models, and the data base underlying them, are seriously deficient for purposes of explaining and predicting expenditures. For this reason, they have also been deficient in attempting to measure the impact of changes in fiscal, monetary, and other economic policies on changes in the spending-saving rate. Since the major portion of aggregate demand and of changes in demand emanate from the consumer sector, this means that microeconomic stabilization policy cannot be accurately and confidently specified at the present time with existing models. In short, we know substantially less about consumer spending-saving behavior than is necessary either to understand or predict consumer behavior with sufficient accuracy to meet public policy needs, especially around the vicinity of turning points in economic activity.

It is interesting to speculate about the reasons for this, in view of the enormous input of highly skilled professional resources into this area. Some of the problems with the aggregate model approach are reasonably clear: not all households or groups of households react to a given stimulus in the same way; expectations about future developments clearly play an important role in spending and saving decisions; and the available number of degrees of freedom in aggregate macrodata are simply insufficient to permit a test or even an adequate specification of the appropriate model. To be more specific: it is probably true that income changes mean one thing for young families on the way up and quite another for mature families whose prospects have stabilized and whose stocks of durables are larger; uncertainty about the future or a change in uncertainty may have a quite different effect for families who view their long-term prospects as being highly favorable than for those who view them otherwise; the earnings of secondary workers such as wives and older children may well have an impact on expenditures whether the earnings are real or only potential, since wives who have potentially high market earnings are likely to influence household spending behavior whether or not they are actually in the labor force; perceived needs for retirement income must have been changing substantially during a period when social security benefits, private pension coverage, and current income have all been expanding rapidly; and so on. Aggregate economic life is so highly collinear that it is virtually impossible to disentangle these effects in time-series.

The discussion suggests some of what needs to be done to improve our knowledge of the spending-saving relation among households.

1. There is need to work with data at the micro-level in order to specify more precisely the relations among income, saving and consumption for different groups of households. No present set of microdata provides dynamic information on household income, consumption and saving.

2. The needed microdata must have the characteristics of being relatively free from response and measurement error. One of the major advantages of using aggregate data, and one of the reasons why these are used so often, is that the aggregation or averaging process tends to eliminate response and measurement error. In moving to analysis of micromodels, the response-measurement error problem becomes serious, since sufficient error will tend to introduce unknown (typically downward) biases in the coefficients of any model. Present measurement technology is probably not able to produce good microdata with sufficiently low measurement error to meet policy needs.

3. We need to improve our knowledge of the role of consumer expectations, anticipations and uncertainty on spending-saving behavior. Events over the past decade (e.g., the response in the 1968 surtax and the 1972 overwithholding episode, the high saving rates in 1971-72, the burst of anticipatory buying during the first half of 1973, the impact of the 1975 tax cut and the dramatic decline in saving rates since then) have made it clear that expectational phenomena make a difference to behavior, and that the relevant expectations cannot be accurately predicted from a simple extrapolation of past values of objective variables. In particular, it has become clear that spending-saving behavior is strongly influenced by changing expectations about price inflation.[3]

It is the argument of this paper that improvements in our ability to explain consumer spending and saving behavior cannot be accomplished without a basic change in the accuracy with which micromodels of consumer behavior can be estimated. Attractive as the time-series approach is, it does not seem plausible that any real improvement in understanding aggregate behavior can be achieved unless we can construct models which reflect microbehavior, which do so on the basis of data with a substantially lower error component than has typically been the case, and which contain the basic dynamic characteristics of measuring changes at a micro-level over time.

The spirit of the paper is that conventional ways of measuring consumption, saving and income for individual households are incapable of yielding the degree of precision required for estimating the kind of micromodel behavior that seems required to advance the state of knowledge in this area. Despite considerable improvement in the technology of measurement, we are still in a position where basic reliance is placed on the recollection of houshold respondents about the financial flows within the household sector. In the case of expenditure surveys, recollection is, where possible, aided by records relating to the transaction, but there has not been any systematic attempt to base the collection of data on records rather than recall.

The basic argument is that the collection of household sector financial flow data based entirely or mainly on financial records is a feasible undertaking, and that this type of microdata represents one possible solution to the problem of reducing measurement error to the point where micromodels are not dominated by the error component in both dependent and independent variables. The paper also takes the view that appropriate micromodels not only need highly accurate measurements of the relevant financial flows in order to understand spending and saving behavior, but that these data need to be supplemented with a range of anticipatory data--expectations, plans and attitudes, which are in general not uniquely related to past behavior, but rather reflect partly unknown combinations of variables.

Availability of Financial Records

In an economy where financial records were either nonexistent or not systematically maintained, basing an expenditure, income and saving survey on such records would be useless. But in an economy like the U. S., an extremely large fraction of all household financial flows are obtainable directly from easily accessible records. This statement is more true today than it was a decade ago, and more true then than two decades ago. The basic financial record for most households is a checkbook--more precisely the checkbook stub in which expenditures and receipts are recorded. According to a 1973 pilot study conducted at the Survey Research Center, approximately 85 percent of all U. S. households have checking accounts. The proportion of financial flows that go through checking accounts is larger than that, because households who lack checking accounts are apt to be poorer, older, and less well educated than the population at large.

Needless to say, not all these flows can be attributed to specific kinds of expenditures, since many checking account entries reflect cash withdrawals in one form or other. These proportions also vary considerably by income, age and education of respondent as indicated in Table 1. For households with incomes in excess of $15,000, the fraction of total expenditures represented by cash outlays tends to be under 40 percent and gets down as low as 30 percent for households with incomes over $35,000 per annum. The proportion of cash outlays also varies with age, being relatively lower for younger respondents than for older ones. This is presumably a generation effect rather than a disguised income effect, since both young and old heads of households tend to have low family income, with the highest income levels being found in the middle age ranges. By education level, there are marked differences in the fraction of outlays for cash, which run about 70 percent for those with less than eight years of schooling to about 35 percent plus for those with sixteen or more years of schooling.

The pilot study also obtained data on the frequency with which people are paid (essential for identifying receipts from records), and the mode of payment (essential for being able to identify whether or not we have accounted for a receipt in a checking account record). For example, Table 2 indicates

TABLE 1

Proportion of Expenditures by Three Types of Payment, by Income, Age and Education

	Cash	Check	Charge
All	.499	.424	.077
Income			
Under $5,000	.615	.341	.044
$ 5,000 - 7,499	.583	.363	.054
$ 7,500 - 9,999	.515	.412	.072
$10,000 - 12,499	.489	.444	.067
$12,500 - 14,999	.449	.467	.083
$15,000 - 17,499	.356	.529	.114
$17,500 - 19,999	.451	.458	.091
$20,000 - 22,499	.414	.525	.060
$22,500 - 24,999	.355	.481	.164
$25,000 - 29,999	.380	.492	.128
$30,000 - 34,999	.338	.514	.148
$35,000 and over	.310	.506	.184
Age of Respondent			
18 - 24	.416	.449	.075
25 - 34	.423	.497	.080
35 - 44	.452	.463	.085
45 - 54	.517	.392	.092
55 - 64	.576	.352	.072
65 and over	.587	.362	.051
Education			
Under 8 years	.700	.272	.028
9 - 11 years	.654	.305	.040
12 years	.509	.428	.063
13 - 15 years	.416	.496	.087
16 years and more	.366	.497	.137

TABLE 2

Basic Financial Records Data for Job Holders (Percent of Families)

	Checking Account Families	No Checking Account Families
A: HOURS WORKED PER WEEK		
< 5	1.5%	0.0%
5 - 9	1.1	0.0
10 - 19	4.7	2.9
20 - 29	4.7	14.7
30 - 39	15.0	2.9
40	44.5	61.8
>40	28.1	17.6
Other, irregular	0.4	0.0
DK, NA	0.0	0.0
B: FREQUENCY OF PAYMENT ON JOB		
Every week	39.2	82.4
Every 2 weeks (twice a month	38.8	8.8
Once a month	13.6	2.9
Other	7.7	2.9
DK, NA	0.7	2.9
C: METHOD OF PAYMENT ON JOB		
Check or Cash		
Check	94.1	79.4
Cash	4.8	14.7
Both	0.4	2.9
Other	0.0	0.0
DK, NA	0.0	2.9
D: FREQUENCY OF OTHER SELECTED RECEIPTS		
Social Security	21.5	29.2
Pension	11.4	7.7
ADC, welfare	3.4	10.8
Income tax refund	62.8	36.9
Insurance dividend	21.2	10.8
Insurance claim	16.9	1.5
Consultation fees	2.1	0.0
Sale of financial assets	11.2	0.0
Dividend from stock, mutual funds	24.0	1.5
Interest from bonds	13.2	3.1
Tips	3.2	0.0
Odd jobs	14.2	12.3
Sale of assets	5.0	1.5
Gifts, bequests	7.5	3.1
Rents from real estate	13.0	3.1
Repayment of debts	5.0	0.0
Land contract, mortgage	3.4	0.0

that 40 percent of household heads are paid every week, just about 40 percent every two weeks, and the remaining 20 percent once a month or less frequently. Of these receipts, almost 95 percent are in the form of a check. Other forms of receipts were also identified with respect to frequency among families: for example, over 20 percent of household heads report a social security receipt in the family unit; over 11 percent a pension receipt; just over 3 percent ADC or welfare payments; almost 63 percent an income tax refund; over 20 percent insurance dividends, almost 17 percent insurance claims, and about 11 percent receipts from the sale of financial assets.

In addition to checking account records, which would be basic instruments involved in a financial flow survey because they represent a continuous and often comprehensive record of transactions, there are also substantially better annual records available than in the past for a number of critical financial flows. Much of the gain here during recent years is a consequence of changes in the reporting requirements associated with the federal income tax. For example, not only do hoholds receive W-2 forms reflecting annual wage payments, but a wide variety of information returns on other types of income are also provided to households by institutions, along with a few expenditure reports (e.g., interest payments, property taxes, and mortgage repayments). In short, the typical U. S. household is in a situation where, without any great effort on their part, the basic ingredients for a comprehensive survey of income, expenditures and saving could be conducted almost entirely on the basis of financial records containing a high degree o precision, many of which are inherently dynamic in the sense that movements over time are automatically reflected.

Although it is easy to demonstrate that the requisite financial records exist in a large fraction of U. S. households, and that they are capable of reflecting an even larger fraction of total financial flows, that does not prove that a survey based on such records is a feasible undertaking. An obvious issue is: how does one obtain the cooperation of households in attempting to conduct such a survey, given the sensitivity of many households to incursions into their private affairs? I do not think the question can be answered a priori, but what indirect evidence we have on related issues suggests that sensitivity in this and other areas is in fact limited to a relatively small fraction of total households and does not necessarily constitute a serious problem.

For example, the consumer expenditure surveys conducted decennially by the U. S. Bureau of Census constitute a major intrusion not only in the household's privacy but also on its time: households are asked to produce detailed estimates of expenditures for each quarter, are asked to document expenditures by records where possible, and are asked extensive and detailed questions about income. The typical household is asked to spend upwards of 10 hours in talking about the financial details of its expenditures and income with the Census interviewers. Despite what many of us would find to be an inordinate set of demands, response rates in such surveys (1972-73) are typically upwards of 90 percent. That is in part due, of course, to the fact that the work of the Census Bureau is widely regarded as essential; in addition, many respondents may tend to think that they have no legal right of refusal simply because the U. S. Census Bureau is taking the survey. But a better explanation, in my judgment, is that most respondents are extremely cooperative provided they are convinced that the survey is useful, and that many respondents regard the demands made by the survey as an interesting diversion from their daily activities and not as an intrusion.

Other very partial evidence is a bit more pessimistic about the prospects for response rates--more precisely, about the probable extent of wholehearted cooperation from respondents, without which the data base would be unusable. Some years back, the Survey Research Center conducted a small experiment in an attempt to examine the structure of response errors in asset surveys.[4] Respondents were asked to complete forms similar to the capital gains reporting form used for income tax purposes, in an attempt to get accurate information on asset holdings, especially holdings of common stock. While about three-quarters of the sample agreed to participate in the study, only a bit more than half actually completed all of the necessary reporting. Response rate and cooperation experience with the 1962 Federal Reserve Board Survey of Financial Characteristics was better,[5] although again the fact that the interviewing was done by the Census Bureau may mean that the results obtained are misleading as to what is obtainable by other survey organizations.

The Experimental Study

In any event, the question cannot be answered without a serious attempt to obtain such data in the field. An experimental study, now in the planning phase, calls for a number of steps to maximize response rates:

1. Research on the question of interviewer cooperation suggests that data quality is higher and response rates improved if respondents are asked to make a formal commitment to cooperate with the study and to provide complete and accurate data.[6]

2. A number of benefits can be provided to the respondent in return for cooperation with the study. These include payment, where experience suggests that follow-up surveys have higher response rates when payment is used, and provision of an "expenditure pattern" report to the respondent, based on a comparison of expenditure patterns for their household with the average pattern for similar households.

It should be kept in mind that the target population for this study is not the entire population of U. S. households. The use of records will inevitably work poorly for families without checking accounts, and we would not propose to include such families in the panel. We also expect to eliminate families below a certain income cutoff, not because the study could not be conducted but because variation in saving behavior would be uninteresting. A cutoff will also be adopted for high income families with substantial

wealth, where the relevant financial flows, while accessible in principle, are likely to be complex, difficult to obtain, and hard to interpret. Moreover, cooperation rates among wealthy families generally tend to be low for these kinds of data. Thus the survey would be focused on the broad range of families in the lower-middle to upper-middle income groups, a population which accounts for the bulk of total income and total spending, and probably for the bulk of the total time-series variance in spending and saving behavior.

Tentative Schedule of Activities

The first step in creation of the data base is to conduct a survey much like the pilot study discussed above, which simply identifies the kind of financial records that would need to be examined in order to obtain a comprehensive picture of the income and expenditure flow for the household.[7] We envision using a questionnaire schedule much like the 1973 Financial Records one, with a check-box appended on which the interviewer would note the types of records that would be needed for that particular household. The initial interview would contain a final section in which the interviewer explained the next phase to the respondent, attempted to extract a commitment from the respondent to proceed with the study, indicated the benefits that would accrue to the respondent (payment, description of expenditure patterns, etc.), scheduled a follow-up interview, and made clear to the respondent exactly what records should be at hand for purposes of the follow-up interview.

The next phase would involve the follow-up interview where actual data on expenditure and income flows would be obtained. What we have in mind here is an interview schedule which is essentially more like an accounting ledger than a survey. The interviewer would identify the relevant type of record, then ask the respondent to read off the characteristics of the relevant entries. For checking accounts, for example, the respondent would be asked to read off the amount of the check, the date, the general characteristics of the expenditure; for deposits, the data would include the source and the amount; and interviewers would note any supplementary information suggesting that additional records should be examined (i.e., a credit card payment, indicating that actual expenditures need to be obtained from the credit card billing record). In the case of checks drawn to cash, or checks made out for more than the amount of the purchase with the difference being drawn in cash, we would expect to get a general classification from respondents as to the types of purchases involved.

The survey would also attempt to get beginning and ending checking account balances, both for their intrisic interest as well as to check the consistency of the flow data. For other financial records, e.g., savings accounts or credit card billings, the same type of information would be sought--type of financial flow, date of transaction, and beginning and ending balances.

Where a checking account or other financial transaction indicated the existence of an asset, for example, a dividend check, insurance payment, owned business, or an investment account, the interviewer would note the existence of the asset and follow-up questions would be asked about both value and changes in value for the assets in question. Housing transactions are the simplest case in point, where mortgage or property tax payments would be followed by questions about characteristics of the mortgage from which amoritization could be inferred. In cases where the follow-up interview indicated that additional records would contain relevant information, respondents would be asked to locate them; alternatively, the interviewer would try to arrange a second follow-up. In all cases, data from the follow-up interview would be carefully examined by the coding and analyses staff for completeness, indications that other records might be obtainable and would be useful, etc., with a call-back being made in cases where additional information appeared to be obtainable.

Current plans call for attempting to get a 12-month history of financial flows from each household in the sample. To insure that the expenditure and income data can be combined with relevant other data, we plan to interview respondents who have previously been involved in the Survey Research Center's Quarterly Surveys of Consumer Attitudes, which obtain a substantial amount of attitudinal/expectational data. Since we want to relate both expectations and changes in expectations to behavior, we need observations on expectations that precede the expenditure and income data. All respondents will have been asked attitude questions on two prior occasions, the earliest of which would either be simultaneous with or prior to the time span covered by the financial flow data. Finally, this experimental phase of the study will be limited geographically to a collection of states in the North Central and Northeast parts of the country, both to maintain better control over the interview situation and to facilitate interviewer training.

If the data collection project is successful, the resulting data base will be exceptionally rich in several dimensions of consumer behavior that have not been adequately treated in either the theoretical or empirical literature. Specifically, such areas as the time-phase relationships among expenditure categories, the timing relationships among receipts and various expenditures, the effects of anticipated changes in receipts on spending and saving, and interactions between receipts (or other "objective" variables) and attitudes/expectations in the determination of spending decisions, would all be represented in the data base.

Model Testing

While theories of expenditure systems, permanent income, stock adjustment, etc., exist and are well known, they are essentially silent about the dynamic timing of decisions to spend and save. Such theories have rarely been put to the test of a real microdata base with matched observations on all relevant variables. Most of these well-known approaches can be expected to fail in

in important ways when applied to such a data base, or will at the very least, require important additional specification inputs to make them work.

The research process thus involves:

1. Creation of the data base.

2. Application of existing analytical approaches to the data base in a search for basic specifications that seem to work, types of specification changes needed to accommodate existing models to the data, and decision-modes that are markedly different from the ones in standard models--e.g., modes based on threshold effects, discrete decisions, lumpiness, etc.

3. Use of the results of the search process to specify a model or set of models that appear consistent both with theory and empirical regularities.

Although the use of financial records for measuring consumption income, and saving as inputs into behavior modeling is the principal purpose of the experiment, an interesting possible by-product is the potential use of financial records reporting as a more accurate and possibly less expensive mode for the collection of consumption expenditure data. For that use, considerably more information would have to be obtained than the information directly available out of financial records. For example, a survey of consumer expenditures would not be satisfied with a checkbook entry of a particular amount drawn to cash, and it would be necessary to obtain detailed purposes for which the cash withdrawal was used. But one might conceive of starting with a basic survey of financial flows via financial records, then supplementing many of the entries with a follow-up examination of the particular expenditures involved. The same problem would arise for credit card payment entries, where an expenditure analysis would be concerned with purchases as well as with payments. But the basic mode of operation seems feasible, and would quite possibly produce more accurate data for certain types of consumption expenditure and for the aggregate. And it may be no worse for other types of expenditures than the conventional procedures now in use.

1. A 1972 monograph by Thomas Mayer (Permanent Income, Wealth and Consumption, University of California Press, 1973) gives some flavor of the voluminous literature in this area. Several Brookings papers on Economic Activity (Bosworth, Hymans, Taylor, Juster and Wachtel), as well as a 1975 paper (Juster and Taylor) in the American Economic Association Papers and Proceedings volume, contains more recent discussions.

2. The basic work is Milton Friedman's Theory of Consumption Function, Princeton University Press for the National Bureau of Economic Research, 1957.

3. See in particular Juster and Wachtel, "Inflation and the Consumer," Brookings papers on Economic Activity, 1972; Juster, "Savings Behavior, Uncertainty and Price Expectations," in the 21st Conference on Economic Outlook, The University of Michigan, 1973; Juster and Taylor, 1975 Papers and Proceedings of the American Economic Association; and Juster, "Inflation and Consumer Savings Behavior--Some Time-Series and Cross-Section Results," paper presented at the CIRET Conference in 1975.

4. The results of these experiments have never been published. A working paper prepared by Louis Mandell describes the experiment and the results.

5. The nonresponse rate for the 1962 Federal Reserve Board Survey of Financial Characteristics was approximately 27 percent overall: in the highest income group, more than half the sample were nonrespondents. See page 15 in the technical note to the SFCC, published in August 1966 as a Federal Reserve technical paper.

6. Oksenberg, L., Vinokur, A., and Cannell, C. F. The Effects of Commitment to Being a Good Respondent on Interview Performance (a research report). Ann Arbor, Mich.: Survey Research Center, The University of Michigan, 1975. Also a chapter in Experiments in Interviewing Techniques: Field Experiments in Health Reporting, C. F. Cannell, L. Oksenberg, and J. M. Converse (Eds.), in press.

7. The 1973 questionnaire schedule can be obtained by request to the author.

A COMPARISON OF ALTERNATIVE PANEL PROCEDURES FOR OBTAINING HEALTH DATA[1]

Seymour Sudman and Linda Bean Lannom, University of Illinois, Urbana-Champaign

Introduction

The importance of reliable national statistics on the incidence of illnesses and the use of and expenditures for health care has led to the establishment of the Health Interview Survey, which is an integral part of the program of the National Center for Health Statistics, and to continuing studies by the National Center for Health Services Research. These surveys have proved of great importance and have provided much valuable data. However, they have also run into problems that continue to defy solution. A major problem is that these surveys depend on recall for periods of up to a year, even though it is known that substantial recall errors may occur. These errors are basically of two types:

1. Omissions--The respondent omits an illness episode or expenditure entirely. These omissions are not random, but are usually concentrated among short illnesses for which hospitalization was not required, or for routine visits to a physician.
2. Telescoping--The episode is remembered, but there is an error in the date so that the episode is remembered as occurring more recently than it did.

An alternative procedure that may help to solve or reduce some of the problems of health surveys is the use of diaries to obtain health care information. Diaries eliminate or greatly reduce the recall problem, as well as reduce interviewing costs. Diaries may present new problems, however, including level of cooperation, errors in record keeping, and possible conditioning effects. Yet, the diary approach has proven very valuable in other types of surveys, and the possibility that diaries may be equally useful in obtaining health information is sufficiently great to warrant their testing in controlled experiments.

In a study currently in progress at the Survey Research Laboratory, University of Illinois, we are attempting to determine the cost-effectiveness of diaries for obtaining health data from a general population sample. Comparisons are being made between the results obtained from diaries, personal and telephone interviews. The effects of differential diary procedures and compensation are also tested. The analyses will compare levels of cooperation and frequencies of health episodes reported by the various methods and by level of education and previous medical history of respondent households. This paper discusses only the levels of cooperation.

Method

It would be anticipated that households with lower education levels and higher levels of ill- would have the greatest difficulty in keeping diary records as well as recalling medical events. For this reason, a disproportionate stratified sample was selected. Specifically, the following procedure was used:

1. The Survey Research Laboratory screened a probability sample of about 6,000 households in Illinois during January-March, 1976, using phone interviews to obtain information on medical experiences in the previous year as well as other demographic information. The results of the initial screener interview are given in Table 1. It may be seen that screener information was obtained from 5,214 households or 81.1 percent of all contacted households. This level of cooperation is excellent, considering that two-thirds of the population in the State of Illinois is concentrated in the Chicago metropolitan area, where cooperation is usually more difficult to obtain. The reasons for this rate were that the screener questionnaire was carefully pretested three separate times, the interviewers had substantial previous telephone experience and advance post cards were sent to respondents outside the City of Chicago where telephone listings were used. In the City of Chicago, random digit dialing was used since about 40 percent of households have unlisted telephone numbers. Of course, advance postcards could not be sent to these households.

No major efforts were made to convert the refusals or to locate the remaining non-contacts. Past experience would suggest that the cooperation rate might have been increased to nearly 90 percen[illegible]f this had been attempted, but that costs would also have risen sharply. It is important to remember that when cooperation rates are discussed they refer to the households who cooperated on the initial screener. Thus, the approximately seven percent of Illinois households without telephones as well as the non-cooperators on the screener are excluded.

2. From this sample of 5,214, a disproportionate stratified sample of 1,446 households was selected (to obtain a final sample of about 1,200) with the stratifying variables being:
   a. Education of female head of household or spouse of male head;
   b. Level of medical experience in the previous year.

The definitions used for education and incidence of health experience were as follows:

Low education: 11 years or less
High education: 12 years or more

Low health incidence: 14 or fewer total health episodes in the past year and six or fewer times of limited activity and six or fewer times that a hospital was visited by all household members combined.

High health incidence: 15 or more total episodes or 7 or more times of limited activities or 7 or more times that a hospital was visited by all household members combined.

The sample of 5,214 households was distributed as follows:

Stratum	N	Sampling interval
1. Low education, low incidence	460	1.28
2. Low education, high incidence	941	2.61
3. High education, low incidence	2,444	6.79
4. High education, high incidence	1,369	3.80

To allow for possible moves, missing

addresses and other problems unrelated to cooperation, an initial sample selection of 360 from each of the strata was used. This meant that the sampling intervals (and thus the weights) for the four strata were those seen above.

3. An initial interview was conducted with all households which were then randomly assigned to one of the following three treatments:
   a. Three personal interviews at monthly intervals
   b. Recruit to keep a diary of medical experiences for three months with total compensation of $15
   c. Recruit to keep a diary with no compensation

Within a stratum, about 100 households received each treatment.

4. The Survey Research Laboratory attempted procedures for reducing costs with half the households in each treatment method. For the personal interviews, half the households were contacted by phone, rather than face-to-face. For the diary methods, half the households were requested to mail diaries in.

5. SRL attempted to maximize the diary mail in cooperation rates by conducting reminder phone calls to respondents from whom diaries were not received within two weeks of the expected date.

Cooperation by Sample Type

The cooperation rates for the initial interview and for the three months that households were asked to participate are shown in Tables 2-4. The data are first split by sample type since differential sampling rates were used.

It may be seen in Table 2 that the highest cooperation rates are obtained from households with higher levels of education and higher levels of health problems. The lowest cooperation is from households with lower education and lower levels of health problems.

The differences are small and not statistically significant on the initial interview, but become larger during the three months. These differences are highly significant after three months using a chi-square test. Overall, there is a difference of 12 percentage points between the 78 percent cooperation rate of households with more education and more health problems as compared to the 66 percent cooperation rate of those with less education and fewer health problems.

Although these results vary by method as seen below, the effects of sample type are consistent over method. That is, there is little interaction between method and sample type. Again, the reader is reminded that these cooperation rates are based on the sample of households which had already cooperated on an initial screening interview.

The results are as expected for effects of education, but are the opposite of those predicted for levels of health problems. In retrospect, it now appears that those with more health concerns find this study more salient and are more willing to cooperate.

Cooperation by Method and Type

Table 3 presents the key results, household cooperation by method and month, controlling for sample type. The control is necessary since cooperation does vary by sample type and the strata were not selected with equal probability. Nevertheless, the same results are observed for all four sample types.

Three major findings emerge from Table 3:
1. Diary pickup methods obtain levels of cooperation as high as those found for repeated personal and phone interviews.
2. Diary mail in methods are substantially worse in obtaining household cooperation than the other methods.
3. Compensation has no significant effect on cooperation for diary pickup methods, but does have a significant effect for the mail procedures.

Initial Cooperation-There are no significant differences in the cooperation rates on the initial interview by method. Except for those households interviewed by telephone, all initial interviews were conducted face-to-face and were identical, regardless of the method to be used later. In some earlier studies in obtaining food expenditure data by diary methods, there was some evidence that interviewer knowledge of the treatment that households would receive later had an effect on initial cooperation. In this study of health data collection, no such evidence of initial interviewer effects is observed.

There is also no significant difference between the cooperation on the initial interview conducted face-to-face and the cooperation when the interview was conducted by telephone.

Cooperation rates varied from seven to 13 percentage points within types, ranging from lows just under 80 percent to highs in the low 90's, but these differences were not significant on the chi-square tests at the .05 level.

Cooperation on Diary Mail in Procedures-Diary mail in procedures are attractive from a cost standpoint because they eliminate the need for interviewer visits after the initial interview. The results in Table 3 indicate, however, that this reduced cost is at the expense of significantly reduced cooperation rates for every sample type. The highly significant chi-square values observed are due entirely to the diary mail in procedures. Overall, while cooperation after three months was about 80 percent for the other methods, averaging over sample type, it was only 54 percent for the diary mail in procedures.

The results summarized over methods are presented in Tables 4A and B although the significance tests are conducted on the uncombined results of Table 3. Table 4A gives the cooperation rates where each sample type is weighted to account for the differential rates of selection. Table 4B gives the unweighted results that summarize Table 3. Although the results of Table 4A are more exact since they take into account the differential

sampling rates, the differences between the weighted and unweighted summaries are quite small.

Given these results, diary mail in procedures do not appear to be an effective method for collecting health data. It might be possible to improve their efficiency if methods could be developed for quickly sending an interviewer to collect a diary if it were not received in the mail. On the other hand, such combined methods might be more difficult to control and thus less cost-effective than a simple diary pick up or personal interview method.

Cooperation on Diary Pick Up and Personal Methods-There are no significant differences between the cooperation rates for the diary pick up procedures and those using face-to-face or telephone interviewing. As in the work done earlier, almost all of the attrition is in the initial interview for diary keeping. The loss of households is only five percentage points, from 84 percent on the initial interview to 79 percent after three months, for the diary pick up method. A similar drop is observed for the telephone procedures. Households assigned to face-to-face interviews appear less likely to refuse initially, but are slightly more likely to refuse the month one and month two interviews, so that the cooperation after three months is similar to that for the diary and phone methods. Even the earlier differences are not statistically significant.

The low drop out rate after the initial period indicates that the three month record keeping period might be extended with little difficulty. The next extension attempted might be to six consecutive months or to three or four months with an additional data collection period of three or four months one year later, using the same pattern as in the Current Population Survey.

There is no evidence that less educated households with more health problems have any more difficulty with diaries than they do with personal or telephone interviews. The procedure adopted initially was to offer to switch methods for households that refused the assigned method. Only 26 households asked that the method be switched and their subsequent cooperation was lower than for other households. There were substantial problems in keeping the control records straight on these respondents, and in retrospect it is not clear that it was worth the effort. For the less educated households with more health problems, cooperation was lower on the phone than with the diary methods, although the results were not significant.

Effects of Compensation on Cooperation-A surprise in this study was the lack of effect of compensation on cooperation in keeping diaries that were picked up. In earlier health diary research in Marshfield, Wisconsin and Chicago, compensation had improved cooperation by about ten percentage points. Similar results had been observed on consumer expenditure surveys using diaries. In this study there is no evidence of any effect of compensation, either initially or after three months for diaries that are picked up. There is a marginally statistically significant difference of 14 percentage points due to compensation for diaries that are mailed in, but even with compensation, the mail in procedure results in substantially lower levels of cooperation. We can only speculate as to the reasons for the differences. Health topics are obviously more salient than purchasing of low cost food items and keeping health records is an easier task since there are many fewer entries required for most households.

Summary

Looking at cooperation and costs, three of the six methods tested in this study seem inferior to the other three. The two diary mail in procedures, although very inexpensive, are unfortunately far below the other methods in the level of cooperation obtained. The average cooperation after three months is only 54 percent on the mail in procedures, which is only about two thirds the cooperation obtained by the other methods. The diary pick up compensation method is the most expensive and produces no higher cooperation than the diary procedure without compensation.

Of the three remaining methods, telephone procedures are clearly least expensive, face-to-face interviews most expensive, with the uncompensated diary pick up procedure in the middle. Selection between these alternatives depends on the accuracy of reporting, which is now being analyzed.

[1]This research was funded by The National Center for Health Services Research, Grant HS 01869-01.

TABLE 1

SCREENER INTERVIEW RESULTS

	N	%
Total sample	7,956	
Non-housing units	1,524	
Total housing units	6,432	100.0
Completed	5,214	81.1
Refused	892	13.8
Non-contacts or unavailable	326	5.1

TABLE 2

COOPERATION BY SAMPLE TYPE AND MONTH

Sample type		Percent Cooperating			
			Month		
	n	Initial	1	2	3
Low education	675	87.9	75.1	71.1	67.9
Low health experience	338	87.0	73.3	69.2	66.0
High health experience	337	88.7	76.9	73.0	69.7
High education	685	87.9	78.5	75.5	74.0
Low health experience	335	84.5	74.3	71.0	69.6
High health experience	350	91.1	82.6	79.7	78.3
$\chi^2(3)$		7.13	9.9	11.04	16.04
Probability		.07	.025	.01	.001

TABLE 3

COOPERATION BY METHOD, SAMPLE TYPE AND MONTH

Sample type	n	Percent Cooperating			
			Month		
		Initial	1	2	3
Low education, low health experience					
Personal	56	89.3	83.9	78.6	78.6
Phone	66	83.3	69.7	66.7	65.2
Diary pickup-compensation	54	87.0	81.5	79.6	77.8
-no compensation	52	84.6	82.7	78.8	78.8
Diary mail-compensation	59	93.2	62.7	59.3	54.2
-no compensation	51	84.3	60.8	52.9	41.2
$\chi^2(5)$		5.28	14.2	17.47	29.74
Probability		.40	.02	.005	<.001
Low education, high health experience					
Personal	60	91.6	90.0	85.0	83.3
Phone	60	85.0	81.6	81.6	81.6
Diary pickup-compensation	56	80.4	75.0	75.0	75.0
-no compensation	52	90.4	4	84.6	82.7
Diary mail-compensation	53	92.5	69.8	62.3	58.5
-no compensation	56	92.9	53.6	48.2	35.7
$\chi^2(5)$		7.66	30.85	31.25	47.81
Probability		.18	<.001	<.001	<.001
High education, low health experience					
Personal	57	89.5	80.7	77.2	75.4
Phone	62	85.5	82.3	82.3	82.3
Diary pickup-compensation	54	79.6	75.9	74.1	72.2
-no compensation	57	78.9	75.4	75.4	73.7
Diary mail-compensation	51	86.3	70.6	64.7	62.7
-no compensation	54	87.0	59.3	50.0	48.1
$\chi^2(5)$		4.11	10.32	17.87	19.65
Probability		.55	.07	.005	.002
High education, high health experience					
Personal	58	94.8	94.8	91.4	91.4
Phone	68	88.2	85.3	83.8	82.4
Diary pickup-compensation	57	94.7	94.7	94.7	94.7
-no compensation	50	90.0	88.0	86.0	86.0
Diary mail-compensation	60	88.3	73.3	66.7	61.7
-no compensation	57	91.2	59.6	56.1	54.4
$\chi^2(5)$		3.63	38.3	40.13	48.03
Probability		.60	<.001	<.001	<.001

TABLE 4A

COOPERATION BY METHOD AND MONTH
(Weighted)

Method	Percent Cooperating			
		Month		
	Initial	1	2	3
Personal	91.3	86.4	82.5	81.4
Phone	86.0	82.7	81.2	80.6
Diary pickup	84.3	81.7	80.1	79.0
Compensation	84.5	81.3	80.3	79.3
No compensation	84.1	82.2	79.9	78.7
Diary mail	88.8	71.0	58.0	54.0
Compensation	88.6	74.8	64.3	60.9
No compensation	89.0	67.2	51.6	47.1

TABLE 4B

COOPERATION BY METHOD AND MONTH
(Unweighted)

Method		Percent Cooperating			
			Month		
	n	Initial	1	2	3
Personal	256	91.3	87.4	83.1	82.3
Phone	231	85.5	79.7	78.5	77.7
Diary pickup	432	85.6	82.9	81.0	80.1
Compensation	221	85.5	81.9	81.0	80.1
No compensation	211	85.8	83.9	80.1	80.1
Diary mail	441	89.6	63.7	57.6	52.2
Compensation	223	90.1	69.1	63.2	59.2
No compensation	218	89.0	58.3	51.8	45.0

DISCUSSION

Martin David, University of Wisconsin

Disproportionate emphasis will be given to comments on the three papers on the Survey of Consumer Expenditure (CES); the continuing magnitude of our likely expenditures on such surveys and the almost complete inattention that they have received amongst academic statisticians justifies that emphasis. We need a complete, scientific, statistically adequate evaluation of the whole CES design. My comments can be summarized under five headings: No memory, No model, No comment, No dice, and some zip.

No memory. Past work of several of the authors is extremely germane and the CES has been discussed at the 1971 and 1975 ASA meetings. I urge readers to look at that material.

We know from past work that consumer expenditures and savings can not be reconciled with incomes reported. We know that there is differential reporting of information in different categories -- vice and casual expenditures being particularly badly reported. We know that the consumer unit is an artifact of the Bureau of the Census, and that most people can only report expenditure behavior accurately in the areas over which they have control. Pearl remembered these past discoveries and structured his discussion accordingly. It would have been extremely pertinent to the evaluations presented by Dippo to do the same. These considerations imply that we are dealing with a problem in measurement that includes both sampling error and response bias. The conceptually desirable procedure for evaluating the results of the CES would be to appeal to a minimum mean square error criterion (MSE).

Neither of the papers appeal to MSE as a choice criterion. The reason is that the design of the CES anihilates many of the comparisons that one might like: a) Some items are excluded from the diary and included in the interview and conversely; b) even where the same items are measured, the period of measurement may be different, with the result that comparable estimates can not be generated, given the well-known decay curves for recall information.[1] As a result this whole exercise of evaluation of individual items appears to be at sea without a rudder or a paddle -- a combination of Hawthorne effects, telescoping, and respondent fatigue make it unclear whether the diary estimates contain more or less bias than the survey interview estimates.

We do obtain one useful clue to this problem from the Dippo paper. First-day of diary-keeping appears to be biassed upwards by telescoping, and it would appear desirable to incorporate that finding into the estimation procedure used to obtain expenditure aggregates. It is not clear whether that has been done for CPI revision.

However, the Dippo finding is marred by the fact that we learn nothing about the treatment of non-response (including the 10 percent missing data diaries in which interviewer treatment of the first day is not known) in her calculations. How much of an effect does weighting the data have on Table 7?

No model. Both the Pearl and Dippo papers proceed as if we were in a state of ignorance about the nature of response effects and an appropriate psychological model to use for predicting poor performance. Work by Locander, Sudman, and Bradburn,[2] and by Cannell, Oksenberg, and Vinokur[3] give some clues on where and why to expect bias in the use of alternative data collection instruments. Failure to obtain relevant information can either be due to lack of motivation or perceived threats to the respondent from giving the information requested. It would be highly desirable to integrate findings from the evaluation of the CES within this theoretical structure.

The lack of a model, and the lack of emphasis on mean square error minimization imply that Dippo and Pearl reach conflicting conclusions on which data source to use. In large part this is due to the fact that Dippo et al. find significant differences between the diary and the quarterly interview where Pearl reports none. I am astonished to find two users reaching different conclusions or so basic a question. However, it is also the case that Pearl appears to base his choice of the two data collection methods on consistency with the national aggregates (which may themselves not be correct) whereas Dippo et al. use a criterion based on the coefficient of variation (CV). Looking at Pearl's Table 1 does not convince me that the ratio is a compelling criterion for choice -- Is .73 for "clothing" good? Is 1.11 for "food away from home" good? How does this compare to 1961 CES? The nature and logic for a CV choice is also not clear:

a. In the first place a multi-variate procedure would appear desirable for choosing the data collection technique, grouping classes of items together that could be expected to have similar problems in terms of threat, motivation, or recall.

b. Choice of the diary as a preferred source of data when the mean is larger than that for the interview and the CV is not, appears to imply that both numbers are subject only to under-reporting. Therefore larger means represent more complete data, not telescoping, misclassification in the diary or other errors. This assumption should be examined carefully.

The third criticism that I levy under the heading of no model is that both evaluations proceed without reference to the statistical problem for which the CES data were generated, namely revising the weights in the CPI index. It is in the nature of the price index problem that revision is required to maintain a focus on

the quantities that figure importantly in the consumer budget as new products are introduced and relative price changes shift. The design of the SCE must be evaluated by answering the following questions:

a. How does increased disaggregation of products in the weights contribute to the validity of the CPI? Increased disaggregation implies biases due to response effects and burden on the respondent that I feel are unlikely to be compensated by improved validity in the index numbers generated.

b. How does the CES assist in the timely revision of weights? The elapsed time between what is really happening in the world and the capacity of the BLS-Census to update the index makes it hard to believe that the design being evaluated today is reasonable and a cost-effective use of the nation's statistical resources. This is an echo of Pearl's comment on Jacob's paper at the 1975 meeting.

c. Finally, the evaluation of the CES must answer the question -- how do the data collected enhance the capacity of the government to move towards a utility-based cost-of-living price index that reduces the need for revisions of expenditure weights? My own interpretation of the CES is that it does not move us very far in the direction of being able to estimate the systems of structural demand equations that are required for a utility-based index, precisely because the data collection design did not adequately anticipate how to integrate information from the diary and the interview.

No comment. I have briefly mentioned the need for memory. Let me remind Bob Pearl that when he introduced the design for the CES to this association in 1971, he asserted that the design was novel and important because of eight features. His evaluation touches on three of the eight -- quarterly interview vs. diary data collection and the inventory method. Dippo et al. tell us something about the diary-keeping procedure. But several of the features embedded in the design are not touched on in their talks today:

a) Have we learned something about the last payment technique?
b) Has the scheduling of the sample as a time subsample of months and weeks been helpful?
c) How has the awkward problem of migrant families affected the data quality?

In the same meetings in 1971 Lester Frankel commented on the ingenious blending of different samples in the SCE design. 1- and 2-week samples for frequent items of purchase, monthly, quarterly, or annual samples for other items. This complex sample design and the related pattern of recall periods has only been indirectly discussed today, and I feel the profession deserves a report on its strengths and weaknesses. I hope we will see comments on these features in the evaluation reports now being prepared.

Fortunately, we do have some answers today on another feature of the design -- compensation incentives. Cowan's paper gives a clear and admirably documented report on the CES compensation experiment. His paper focuses on reduction in response bias, with the implicit model being a model of respondent omission. His conclusion that compensation is not an important technique for improving response quality must be qualified. The data do point to the fact that increased numbers of responses and response amounts attributable to compensation are a very small fraction of overall variance. What his Eta values do not display is: a. The possible increase in overall response rates that may be associated with compensation. b. Moreover they do not reflect the importance of additional reporting in relation to a measure of mean square error that appears to me to be the appropriate criterion.

Sudman's study also gives us some insight into compensation, and he should be urged to look beyond cooperation rates to the kind of item response analysis that occupied Cowan.

What is interesting about both studies is the light that they shed on the question of respondent motivation. Cannell and his coworkers have found that making the reporting task relevant to the respondent and educating him as to what constitutes a good job is crucial to the complete reporting of health events. Sudman's data demonstrate this effect in the lower cooperation levels of those who have few health events to discuss. The same framework suggests that money should be a more significant motivator to those for whom the task is relevant (i.e., health events to report and for whom incomes are low. This appears to be borne out by Sudman's Table 3 for the mail returned Diary.

Cowan's Table 3 has the potential for giving similar insights, when we see the direction of the interaction effects, which ought to be included. The fact that the interaction effects are strong for urbanicity, race and education offers the possibility that the sensible use of compensation is not to offer compensation to all respondents but to adopt a selective strategy. Identify those in the population for whom money is a good motivator and who are lacking in motivation; then concentrate payments on those individuals. It would seem quite feasible to concentrate compensation, say on urban blacks, if the interaction effects suggest that response in the sub-group could be substantially improved.

No dice. I said at the outset that my penultimate comment was no dice. I refer to the continuing consumer expenditure survey. The inconclusive character of the evaluation of survey interview versus diary data that is reflected in the papers here today stems from a fundamental lack of integration between the

theory of price indices and the measurement processes. This lack of integration was compounded in the CES by the use of independent diary and interview samples. Design of future collections should proceed with a more integrated approach based on the theory of utility-based cost of living price indexes. That will require that diaries and expenditure data be collected from the same sample, together with price statistics so that the behavioral response of consumers to a changing environment can be modelled. Such a design could be carried out with the resources that are required for a CCES. Pearl's recommendations to retain analytical opportunities are particularly wise when viewed from this perspective.

I strongly urge BLS-Census to involve academic statisticians in the design of CCES so that the product can be more useful than the piecemeal CES.

Some zip. My last comment is that Juster's proposal has some zip. The practitioners in the field of expenditure measurement appear to have forgotten that we live in a space age in which technology can be used to assist in data collection. The profession should not be devising ways of burdening the respondent with more forms paperwork or hours of interview -- we should be devising ways of automatically recording behavior as it occurs. The Neilson ratings do this. A pocket electronic memory could be devised that might substantially increase the coverage and accuracy of diary methods.

Juster's suggestion that we look at checkbook records is another way of automating data collection. It probably ought to be supplemented with data from credit card statements, and I am sure that a checkbook study ought at the beginning to be done on a very limited time scale such as the 2-week diaries we have heard about today. I also would caution that the reconcilliation of records with the social scientist's conceptual structure of income and savings is extremely difficult. Tax records are not economic records. Finally Juster implies that we must know the inventory or cash equivalents which are among the most difficult data to get completely reported. Beyond that I can only say good luck!

---

1. Norman Bradburn and Seymour Sudman, Response Effects in Surveys (1974).
2. ASA Proceedings: Social Statistics Section 1975.
3. Journal of Marketing (1977).

MEASURING CRIME BY MAIL SURVEYS: THE TEXAS CRIME TREND SURVEY

Alfred St. Louis, Texas Department of Public Safety

Introduction

The Texas Crime Trend Survey is a mail survey of the general public. The purpose of the Survey is the measurement of crime and the level of reporting of crime by citizens to the police. In addition to measuring levels of crime and reporting, the attitudes and expectations of the public are also queried. The results of the survey are widely distributed to criminal justice agency administrators and planners, and also to the general public through the press. The survey is a new crime information system based upon the reports of crime victims and the general public.

Sample

The Texas Crime Trend Survey was initiated in March, 1976. The original design calls for the survey to be conducted every six months, in January and July. The sample size is 1000, and the sample is a systematic random sample drawn from the computerized Texas Drivers License file which is maintained by the Texas Department of Public Safety. The focus of the survey is the individual driver's experience with crime, rather than the household unit. The survey is not a panel study, and new names are surveyed every six months. While the survey of 1000 people is repeated every six months, the length of the reference period is 12 months. Each respondent is queried about his or her experience with crime during the 12 months prior to the survey. Thus, each successive survey covers 6 months that were previously covered, as well as the most recent 6 month period. The effect of the 12 month reference period is continuously overlapping surveys. The result of overlapping reference periods is that data from 2000 people are available for analysis when 2 surveys with overlapping time periods are combined. By utilizing the technique of overlapping time periods the 2 sets of data for each time period can be compared to each other for purposes of validating the measures of crime.

Methodology

The survey is conducted by mail using a visually attractive booklet questionnaire illustrated with cartoons. The methodology is based on the work of sociologist Don Dillman and his colleagues at the University of Washington.[1] The main principle is persistent follow-up. The persistent follow-ups overcome the most serious shortcomings of mail surveys - the generally low response rates. When response rates are below 50%, and this is common when extensive follow-ups are not used, the data are of limited value in providing accurate estimates. The methodology used in the Texas Crime Trend Survey has consistently produced response rates between 84% and 86%.

The procedure used to contact people in the sample begins with a cover letter and clearly numbered questionnaire. After 2 weeks a follow-up postcard is mailed to non-respondents. The initial mailing and one follow-up produces about 60% of the sample. After 4 weeks from the initial mailing a second cover letter and questionnaire are mailed to about 350-400 people who have not responded. About half will respond, and the other half are mailed a 2nd postcard 6 weeks from the initial mailing. After 8 weeks the response rate averages between 84 and 86%.

The remainder of the sample is then telephoned to estimate the non-response effects. Only half of the non-respondents can be reached by phone, because they either do not have one, or they have unlisted numbers, or have moved, died, etc. Of the people who do have accessible phone numbers, half are successfully interviewed to estimate non-response effects. The telephone follow-up stimulates more questionnaire returns, but they are usually too late to include in the analysis. Generally, the bias in the response rate is in the direction of prior victimization. The people who have been victims are more likely to return the booklet promptly.[2] Thus, victimization implies interest and greater motivation to participate in the survey. The response rate was 84.4% for the first survey and 84.7% for the second survey.

Beginning with the third survey a Spanish translation of the questionnaire was mailed to all persons with a Spanish surname who were non-respondents at the time of the 2nd follow-up mailing. This translation increased the response rate to 85.6%. Several additional factors which are operating to produce the high response rate are the legitimacy of the agency conducting the survey, the Texas Department of Public Safety, and the public interest in the topic. The Texas Department of Public Safety includes the Highway Patrol, Disaster Emergency Services, and the Criminal Investigation Division. The good public image and the professionalism of the Department employees is recognized throughout the state. Also, increasing crime rates in Texas have contributed to increased public interest in the topic. The 1977 session of the Texas Legislature included a widely publicized package of bills aimed at "Crime Control".

Texas Crime Victim Index

The data collected from the Texas Crime Trend Surveys are used to develop the Texas Crime Victim Index. This Index measures the percentage of the population who are victims of crime. The Index is analogous to the IACP-FBI's Index of Serious Crime which is popularly known as the crime rate. However, while the FBI Index is presented in crime events per 100,000 population, the Texas Crime Victim Index uses the person as the unit of analysis rather than the crime event. In the Victim Index if 20 people out of 100 experienced 30 crimes in the past year, the result would be an index of 20 percent. The FBI Index, based on crime events, would score this as a rate of 30,000 per 100,000.[3]

The purpose of developing the Crime Victim Index and presenting it in a simple percentage format is to improve public understanding of the crime rate and the risk of crime. This emphasis on the communicability of crime statistics has been recommended by the recent report of the National Academy of Sciences: Surveying Crime.[4] The presentation and display of crime data to the public in an easily understood format should enable people to assess their personal exposure and vulnerability to crime, and to react accordingly. Just as we are now being told by public health officials that the next great advances in the longevity of life will have to come from the individual's own efforts to respond and react to his environment, the same principle may be applied to crime control. The efforts of individual citizens to reduce their exposure to the risk of crime is a promising area of future research in crime prevention and control. Comments and letters received from survey respondents indicate that some people are acutely aware of this approach, and have already reacted by taking measures to reduce their risk of both property and violent crime.

The data from 3 surveys have been analyzed, and trends have been developed. The Texas Crime Victim Index registered a statistically significant increase in 1976 when compared to the 1975 baseline data. The percentage of victims in the population increased from 17.9% to 21.6%. The definition of victim is operationally defined by the responses to the seven types of crime queried in the survey booklet: Burglary, Robbery, Rape, Assault with Weapon, Assault with Body, Motor Vehicle Theft, and Other Theft. If a person reported they were a victim of one or more of these crimes then the computer program classified them as a victim. Attempts were classified separately from victims for quality control purposes. Because the survey is involved in measuring crime as perceived by the respondent, some attempted crimes could easily be dismissed as projections of the imagination. Therefore, to insure a stringent definition of crimes reported attempts are classified and analyzed separately from completed crimes.

The Texas Crime Victim Index is divided into the Violent Crime Victim Index and the Property Crime Victim Index. The 1976 indices registered a 5.2% violence index and a 16.4% property index. The violence index is composed of Robbery, Rape, and Assaults. The property index is composed of Burglary, Motor Vehicle Theft, and Other Thefts. Both indices are composite indices, and the separate crime types are unweighted. Theft accounts for most of the crime events in the raw data used to construct the Texas Crime Victim Index, followed by Burglary which is the second most frequent of the crime events. Therefore, the Texas Crime Victim Index shares the same characteristics of the weighting problem as the unweighted FBI index of crime. The components of both indices are unweighted, and each of the composite indices is strongly influenced by Theft which is the most frequent crime. However, even though the Victim Index is unweighted, it may be a more sensitive measure of violence than the FBI index.

Index Analysis

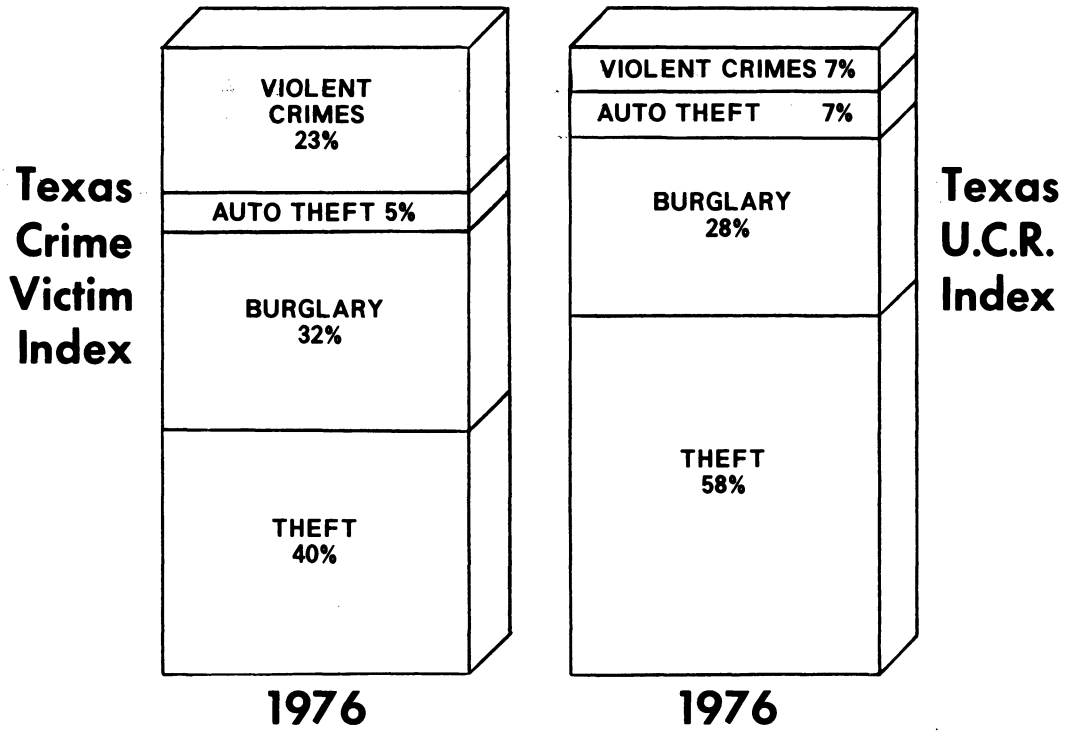
The FBI index of crime, the Uniform Crime Reports, indicates that violence accounts for about 7% of all 1976 Texas crime included in the index, and the remaining 93% is classified as property crime.[5] In the Texas Crime Victim Index violence accounts for 23% of the total 1976 index, while property crime accounts for the remaining 77%. The two indices are compared in Graph A. The definitions of violence differ in the Uniform Crime Reports and the Texas Crime Trend Survey, so direct comparison of the distinct measures is at best speculative, but it is used here for heuristic purposes. The main differences in the data collection systems between the FBI Index and the Texas Crime Trend Survey have been previously acknowledged and summarized elsewhere.[6]

The violence index used by the FBI includes Homicide, Aggravated Assault, Robbery, and Rape. The Texas Violent Crime Victim Index does not include Homicide, includes only completed rapes, and includes assaults that do not meet the FBI's definitional requirements of "aggravated assault". The most frequent crime of the violent crimes queried in the Texas Crime Trend Survey is Assault with Body. No doubt many of these assaults would probably be classified as "simple assaults" according to the definitions contained in the Uniform Crime Report guidelines. However, since the Uniform Crime Reports Index contains many petty thefts, especially since 1972 when the $50 minimum on crimes of theft was dropped, it could be argued that the Index is overly weighted by petty thefts, and underweighted by violence such as assaults which do not meet the strict definition of aggravated.

The unweighted Index of the Uniform Crime Reports for Texas is increasingly dominated by the crime of Theft. In 1970, the crime of Theft accounted for 30% of the Index crimes. By 1976 the crime of Theft accounted for almost 59% of the Index Crimes. During the same six year period the 4 violent crimes share of the Index decreased from 13% to less than 7% of the Index. The internal changes in the unweighted Index, namely the dropping of the $50 minimum value of Thefts, have produced serious change in the FBI's Index of Serious Crime: the Index is being dominated by the least serious of the seven crime types. Projecting into the future, if this trend continues the composition of the crime index in 1980 will be 75% Theft, and 25% for the other six crime types. One way to overcome this continuing trend is to include less serious crimes of violence in the Index. The Texas Crime Trend Survey includes the crime of Assault in its Index of crime, and the result is an index that is not quite so dominated by the crime of theft. However, more data will have to be collected to insure the reliability of the Texas Violent Crime Index. Also, the Texas Crime Victim Index still shares major characteristics of the FBI Index: both are unweighted by crime type, and both include petty theft.

There have been numerous critiques of the disadvantages of an unweighted crime index, as well as several major efforts to weight the

Graph A MEASURES OF CRIME



individual component crimes of the index. Despite the conceptual disadvantages of unweighted indices of crime it is difficult to improve them. Blumstein's analysis of the FBI Index concluded that attempts to weight the individual crimes in the index did not appreciably add to the information communicated in the Index over time.[7] The implications of Blumstein's analysis are: (1) leave the FBI Index unweighted as is, and (2) develop other indices to measure specific crimes or groups of crimes. Therefore, there is a need for multiple indices of crime, but do not change the FBI Index because it works well as designed. The purpose of developing the Texas Crime Victim Index is to complement the information available from the IACP-FBI Index.

The Texas Violent Crime Index increased from 4.2% in 1975 to 5.2% in 1976, but the difference was not statistically significant at the .05 level. This means that the percentage of the population who were victims of violence in 1976 was estimated to be 5.2%. This comparison was made with sample sizes of 1000 and in future comparisons when samples of 2000 are available the possibility of statistically significant results will be enhanced by the larger N's.

The change in the Property Crime Index between 1975 and 1976 was statistically significant, from 13.7% to 16.4% of the population. The sample sizes of 1000 were sufficient to detect the change in property crime at the .05 level. The next report comparing two complete years of data from the surveys, the comparison of 75-76 with 76-77, will have sample sizes of 2000 for all time periods when the data from 1977 is collected in February, 1978.

Trend Data

The data on trends over time in the Crime Victim Index can be presented by month of occurrence or any time period less than 1 year, as the month is queried in the survey. When the data are displayed in six month periods, which is a convenient time frame because of the semi-annual data collection and overlapping reference periods, the results of successive surveys can be combined. The two year trend displayed in the Texas Crime Victim Index is generally stable with the exception of the first 6 months of the data, the January to June period of 1975. The Victim Index for successive six month periods was: 14.4%, 21.8%, 21.0%, 21.9%. The second and third percentages are averages of two samples combined, and therefore represent a total sample size of 2000. The first and fourth percentages are based on only one sample of 1000 each. The fourth percentage, 21.9%, will be averaged with data from the current survey which also covers the last six months of 1976, as well as the first six months of 1977.

The anomaly in the two year trend is the first six months of data collected, the January to June, 1975 data. The low index level, 14.4%, could have occurred because the first survey was almost three months behind the mailing schedule. Instead of being mailed on January 1, 1976, the survey was mailed on March 20, 1976. The result

was that a reference period of 15 months was used instead of 12 months as originally planned. The effect of this lengthened reference period could be the cause of the relatively low level of crime measured for early 1975. A longer reference period implies memory decay, and some previous research conducted by Biderman suggests that memory loss is a critical variable.[8] Fortunately, for Index development purposes the first six months of data can be dropped from consideration because only one sample of data is available for that time period. The accuracy of the Crime Victim Index is improved by utilizing only time periods covered by two overlapping samples. These double measures of the crime level will be useful in detecting extreme variation in trends.

The accuracy of the Texas Crime Victim Index has yet to be conclusively demonstrated as it is in a developmental stage of growth and increases in the sample size are planned. However, there is some evidence that the Index will be reasonably accurate when the developmental efforts are completed. The two time periods that were covered by successive samples were the last six months of 1975 and the first six months of 1976. The Index measure was within 1% for each of the two separate periods. In the second half of 1975 the two separate samples measured the crime level at 21.4% and 22.1%, a difference of only .7%. For the first half of 1976 the two separate samples measured the crime level at 21.3% and 20.7%, a difference of .6%. The standard error is 1.2% for a 1000 sample size, so both of these tests were well within the standard error. This demonstration of the accuracy of the Index is not conclusive proof, but it is encouraging information suggesting that further investment in this Index development will have a high probability of success. The cost of conducting the Texas Crime Trend Survey even with an expanded sample size will be less than the cost of any other comparable measure of crime. However, the accuracy of the measurement of crime levels in society is a subject worthy of at least two separate and distinct indicators. Both the Uniform Crime Reports and some measure of the Victim experience such as the Texas Crime Victim Index should be continuously refined to monitor the crime rate. The cost of the criminal justice system in Texas is rapidly approaching $1 billion annually, and this expenditure alone is sufficient to justify investing in accurate measures of crime.

The violence index is not nearly as stable as the Texas Crime Victim Index. The percentages of the two separate measures for the last half of 1975 were 5.0% and 7.0%. For the first 6 months of 1976 the two measures were 6.6% and 5.2%. The standard error is .7% for the violence index, and this value was exceeded in the 1975 measures. Because the violence measure is a relatively small part of the sample the accuracy is expectedly lower, and therefore less stable. Larger samples will be necessary to develop accurate measures of violence. Future plans for the survey include increasing the sample size to 4000 or 5000 per survey. The goal of the survey operation is to continue to keep the costs low while automating as much of the mailing and data processing without losing the personalized letter format. Until some technical problems in automating the data collection are solved the sample size will not be increased.

Costs

The cost of conducting the Texas Crime Trend Survey is estimated at $3 per completed survey booklet. This cost is very low compared to other data collection methods. A recent study estimated the costs of conducting crime surveys by telephone interviews around $30 per interview, while the current LEAA - Bureau of the Census personal, face-to-face interviews were estimated to cost $100 per interview.[9] Traditionally, mail has always been viewed as the cheapest method of collecting data. The low response rates from mail surveys have prompted more expensive personal interviews. But, if the public is interested in the topic, as is the case with the topic of crime, and good follow-up techniques are utilized, then the non-response problem is effectively solved, and costs are kept low. Mail collectic aves by transfering the labor costs from the interviewer to the interviewee. This savings in labor is partially offset by the disadvantage of one-time feedback from the respondent. No clarification can be made on ambiguous responses. However, since 75 to 80% of the sample are non-victims during the 12 months reference period, the ambiguous responses apply to only a fraction of the total sample. To be sure, the victims of crime are a small fraction of the total sample, but they are the most important part of the sample in terms of the analysis of the data. Therefore, any techniques to reduce ambiguities in the questions and responses will help insure accurate measurement.

Comparison of Results

The data collected by mail have been compared to other data bases of crime data including the FBI Uniform Crime Reports for Texas, 1975, and the Texas Department of Public Safety UCR program, 1976. The crime survey data are not directly comparable to the UCR because the definitions of crime differ. However, the overall pattern of crime uncovered by survey is similar to the pattern of crime reported by police. Most of the crime measured by both of these methods is Theft, followed by Burglary which is second in volume. There are differences between the volumes due to reporting and non-reporting, but the iceberg theory does not hold. That is, reported crime is not the proverbial tip of the iceberg, as the most serious crime is reported to the police. The reporting of crime varies directly with seriousness of the crime, both in terms of violence and dollar loss amounts. The bulk of unreported crime is thefts with small losses, under $200. The crimes that are unreported vary by crime type, but, generally the picture of crime portrayed by data from victim surveys is very similar to the pattern in the Uniform Crime

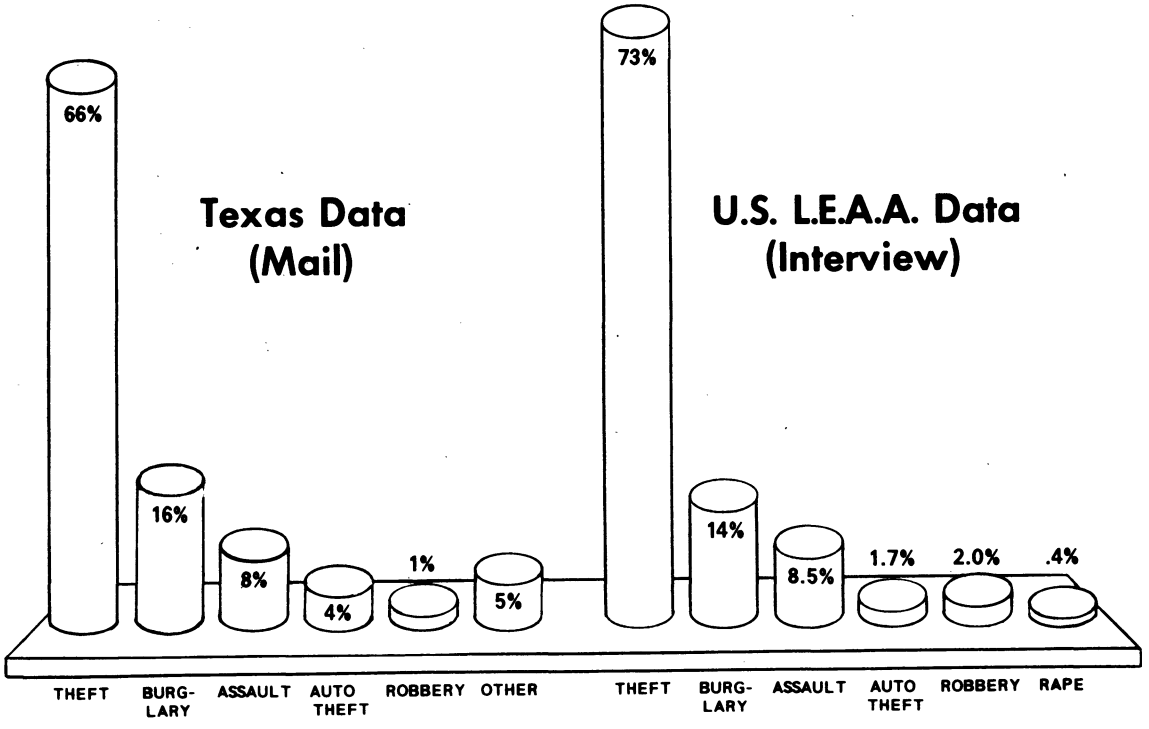
Reports.

The unreported crime data from the Texas Crime Trend Survey have also been compared to the published data from the National Crime Panel Victimization Survey conducted by the Bureau of the Census under contract with LEAA. The data presented in the LEAA publications are not directly comparable because of different wording in questions, and also because the rates of crime are presented in terms of crime events per 1000 population. The Texas Crime Trend Survey data are presented with the victim as the unit of analysis rather than the crime event. However, some data from the National Crime Panel have been tabulated and published in a format comparable to a breakdown of the Texas data. The data on Unreported Crime Incidents published by Skogan indicate that of all unreported crime in the US in 1973, Larceny-Theft comprised 73% of the total, followed by Burglary with 14%, and Assault with 8.5%.[10] The data from the Texas Crime Trend Survey for 1975 indicate that Theft comprised 66% of all unreported crime events, Burglary 16%, and Assault 8%. The two sets of data, the National for 1973, and the Texas for 1975, are not identical. However, the pattern of unreported crime in both sets of data is very similar, and is illustrated in Graph B. Theft is the most frequent unreported crime, followed by Burglary, Assault, etc. This similarity of patterns indicates that the measurement of unreported crime is reasonably consistent, even when different methods of collection, mail and personal interview, are used. Regardless of the data collection method the general pattern of unreported crime is consistent. This is not an attempt to ignore the real differences involved in different methods of collection, but simply an effort to illustrate the reliability of data collected by mail questionnaire. For example, more methodological research would be required to see which method, mail or personal interview, is more likely to elicit information from rape victims. While mail and personal interview methods produce a similar general pattern of unreported crime, there may be specific areas of systematic variation associated with each different method whether it be telephone, mail or personal interview. Herman found that telephone interviews may not be as good as personal interviews for sensitive data such as illegal behavior or voting decisions.[11]

Additional Survey Findings

In addition to measuring the level of crime in the state, the survey measures the level of reporting and non-reporting to the police. This information is of particular value to the police. The reasons for non-reporting have also been analyzed, and the main conclusion is that reporting is primarily a function of the seriousness of the crime event. The more serious or costly a crime is to the victim, the more likely it will be reported to the police. There are variations in

Graph B PATTERN OF UNREPORTED CRIME



reporting by type of crime, however, especially regarding rapes and attempted rapes where embarrassment and stigma reduce reporting levels.

Also included in the survey reports are data on losses due to crime. In 1975 the average loss per adult Texan was estimated to be $98. In 1976 the average loss increased to $109 per adult Texan. The expectations of future crime are queried in the survey, and there was a slight increase in the fear of crime for 1976. The victims fear of crime increased from 31% to 33% in 1976. This means that one-third of the victims expected to be victimized again in 1977. The fear of crime among non-victims is lower, only 14% of the 1976 non-victims expected a crime in 1977. The expectations of the public regarding crime are potentially a sensitive measure of future crime events, as well as a measure of the general fear of crime.

Other data available from the survey are the rural-urban distribution of crime, the risk of crime by age, sex, race and ethnic background of the survey respondents, income levels and risk, etc. Relationships among these variables have been summarized in previous survey publications. In brief, there are many possibilities for new analyses of data that have previously been unobtainable because of the lack of an information system focusing on the general public and the crime victim.

References

1. Dillman, Don A., Christenson, James A., Carpenter, Edwin H., Brooks, Ralph M., Increasing Mail Questionnnaire Response: A Four State Comparison. American Sociological Review, October, 1974. pp. 744-756.
2. Sen, A. R., Developments In Migratory Game Bird Surveys, Journal of The American Statistical Association, March, 1976, Volume 71, Number 353. pp. 43-48.
3. U. S. Department of Justice, Federal Bureau of Investigation, Uniform Crime Reports, Crime in the United States 1975, August, 1976.
4. National Academy of Sciences (Panel for the Evaluation of Crime Surveys) Surveying Crime. Washington D.C. 1976.
5. Texas Department of Public Safety, Crime In Texas, 1976.
6. St. Louis, Alfred, Victim Reports of Crime: The 1975-76 Texas Crime Trend Survey. Statistical Analysis Center, Texas Department of Public Safety, Austin, Texas. April, 1977.
7. Blumstein, Alfred, Seriousness Weights in an Index of Crime, American Sociological Review, December, 1974. pp. 854-64.
8. Biderman, Albert D., Surveys of Population Samples for Estimating Crime Incidence. The Annals of the American Academy of Political and Social Science. 374 (1967). pp. 16-33.
9. Tuchfarber, Alfred and Klecka, William R., Random Digit Dialing: Lowering The Cost of Victimization Surveys, The Police Foundation, 1976.
10. Skogan, Wesley G., Dimensions of the Dark Figure of Unreported Crime. Crime and Delinquency, January, 1977. pp. 41-50.
11. Herman, Jeanne Brett, Mixed-Mode Data Collection: Telephone and Personal Interviewing. Journal of Applied Psychology, Forthcoming, 1977.

Footnote

* This research was supported by Grant No. AC-76-F01-4362.

SAMPLE SELECTION PROCEDURES FOR THE IACP UNIFORM CRIME REPORT AUDIT

David W. Chapman*, Westat, Inc.

1. Introduction

In 1974 the International Association of Chiefs of Police (IACP) began a project, sponsored by the Law Enforcement Assistance Administration, to investigate the quality of crime incidence reports that are submitted to the FBI by the Nation's police departments. This project, referred to as the IACP-UCR Audit/Evaluation Project, involves an audit of the processing of various types of information by police departments.

Twenty departments were audited during Phases II and III of the IACP-UCR Audit/Evaluation Project. Since there was no intent to make inferences from the 20 sample departments to all departments in the country, it was not necessary to select the sample on a probability basis. Consequently, the 20 departments were selected on a subjective basis. These test agencies were chosen by IACP personnel to be representative of the police departments across the country with respect to several characteristics.

For audit purposes the processing operation has been broken down into the following four stages:

Stage I - Telephone Tapes (Complaints)
Stage II - Complaint Control Cards
Stage III - Incident/Offense Reports
Stage IV - Clearance Data

Ideally it would be best to audit an agency by checking the accuracy of processing every piece of information at each of the four stages. However, this would be much too expensive and time consuming to do, especially in large departments. Therefore, a procedure was developed to sample the processing of information at the four stages for the audit check.

2. Sample Sizes and the Basic Selection Procedures

An initial decision had to be made between two possible basic selection procedures: (1) independent selection of cases at the four stages, and (2) selection of a sample of cases at Stage I to trace through the system.

Although it might have been useful to trace the processing of cases through the system, there would be a fundamental problem with this procedure. In order to have an adequate sample size for the latter stages (III and IV), a very large sample at Stage I would be required. Since sampling at Stage I (i.e., the telephone tapes) is the most time consuming phase of the audit, this procedure was not used.

Therefore, the first alternative, that of selecting independent samples of cases at each stage, was chosen for the audit procedures. In addition, it was decided to select records in such a way that would provide estimates of processing error rates with the same precision at each of the four stages.

In order to determine adequate sample sizes for selecting cases, the type of estimates to be made and the desired precision of such estimates had to be specified. The basic type of estimate calculated from the audit data is the estimated error rate at one of the four stages. This is defined as the estimated proportion of the cases processed at a stage that is incorrectly classified. The determination of whether or not a case was properly classified was a subjective judgment made by the IACP staff member doing the audit, based on specific guidelines.

For simple random sampling, estimates of the standard error of an estimated error rate can be made using the following well-known formula for the standard error of a sample proportion, p:

$$\sigma_p = \sqrt{\frac{N - n}{N - 1}\frac{PQ}{n}} \qquad (1)$$

where

N = the total number of cases processed at a particular stage during the last month,

n = the sample size,

P = the true error rate for that stage,

Q = 1 - P.

As indicated, the above equation applies to simple random sampling. Actually, systematic random sampling of cases with equal selection probabilities was used for the audit.[1] However, in this situation these two types of sampling procedures probably have about the same precision. For planning purposes the above formula should be adequate to estimate the standard error for an estimated error rate calculated from a systematic random sample.

Based on discussions with IACP personnel, it was agreed that a standard error of .02 for estimating a true rate of .10 and a standard error of .005 for estimating a true error rate of .01 would be adequate precision for the audit estimates. The sample size table that was used most often in the audit procedures, was based on this requirement.[2] This sample size table is Table 1.

3. Selection of the Samples of Cases

For sampling cases at Stages II - IV, the selection procedure was straightforward. The total number of cases processed (i.e., the group size) at each of these stages was usually easy to obtain since these cases were typically listed on cards or records in a file. From the group size, the required sample size was obtained from Table 1. The sample was then selected as a systematic random sample. The selection (or skip) interval used was obtained by dividing the group size (N) by the required sample size (n). (Tables of skip intervals and random digits to select random

*now at U.S. Bureau of the Census

starts were made available to simplify the selection procedures.)

The sampling of the telephone tapes (Stage I) was more complex than was the sampling at the other stages. Very few agencies have a record of the number of calls recorded on their tapes. Even when this is known, the number of these that are relevant to the audit (i.e., that involve at least some minimal crime) is not known.

Therefore, the first step in the sampling of the telephone tapes was to estimate the total number of relevant calls on the tapes for the month. This was done as a two-part procedure. First, the total number of calls in the month was approximated. Next, the ratio of relevant calls to total calls was estimated. From these two quantities an estimate of the total number of relevant cases, N, was calculated.[3] Reference to one of the sample size specification tables (i.e., Tables 1, 2, or 3, depending on the size of the department) provided the target sample size, n, for relevant telephone cases.

For a 30-day month, the number of hours, h, to be monitored was determined by multiplying the sampling rate, n/N, times the total number of hours in the month, 720. It was decided to monitor the tapes in terms of 15-minute segments throughout the month. Therefore, the total number of quarter-hour segments, q, to be monitored was calculated as four times the required number of hours (i.e., q = 4h).

The q segments to be monitored were selected systematically in two stages. First a sample of seven or eight days of the month was obtained by choosing every fourth day of the month, using a random start. (The selection interval of four was chosen to provide coverage of the different days of the week.) The number of 15-minute segments, S, in the days selected was then calculated (i.e., either 7 x 96 or 8 x 96). Finally, the segments to be monitored were selected systematically from the segments in the days chosen. The appropriate selection interval was, of course, S/q. An example of the selection of telephone tape segments is given below.

The calls monitored were all those that originated in any of the 15-minute segments selected for listening.[4] This procedure gave all calls on the tapes for the month an equal chance of selection (i.e., n/N).

The first 20 audits were carried out by IACP personnel with the cooperation and assistance of the police department personnel. It is intended that eventually the audits will be performed entirely by police department personnel. It may be difficult for them to carry out these selection procedures, especially those for the telephone tapes.

Example of the Selection of a Sample of Tape Segments

Estimated total number of calls on tape for February: 32,000

Estimated ratio of total calls to "meaningful" calls: 4:1

Therefore, $N \doteq (.25)(32,000) = 8,000$

Sample size from Table 1: $n = 250$

Sampling rate: $f = 250/8,000 = .01325$

Total number of hours in month: $(28)(24) = 672$

Number of hours to be sampled:
$h = (.01325)(672) = 21$

Number of quarter-hour segments to be sampled:
$q = 4(21) = 84$

Random start for the selection of days: 2

Select systematic random sample of every 4th day beginning with 2nd: 2nd, 6th, 10th, 14th, 18th, 22nd, 26th

Total number of segments in these days:
$S = (7)(96) = 672$

Selection interval for sampling time segments:
672/84 = 8, random start: 3

Obtain sample from time-interval table (Table 4)

Footnotes

[1]This method of selecting cases was chosen since it is a probability sampling procedure that is straightforward enough to eventually be carried out by police department personnel.

[2]Obtaining an adequate number of cases from the telephone tapes (Stage I) was so time consuming for smaller agencies that this precision requirement was relaxed somewhat for Stage I sampling in smaller agencies. The sample size tables used in such cases are Tables 2 and 3.

[3]In some cases the department personnel were not able to provide the estimates needed. In these instances IACP personnel listened to portions of the telephone tapes in order to make these estimates.

[4]In some departments it appeared that as the tape sampling progressed, the total number of meaningful calls selected in the sample segments would differ substantially from the target number. In such cases, the number of sample segments was either increased or decreased in an attempt to bring the sample size close to the target sample size.

Table 1 (.02 Standard Error - True Error Rate of .1 and a .005 Standard Error - True Error Rate of .01)

Group Size	Sample Size
1-60	all
61-80	50
81-120	70
121-200	90
201-500	120
501-1000	200
1,001-Over	250

Table 2 (.025 Standard Error - True Error Rate of .1)

Group Size	Sample Size
1-60	all
61-80	50
81-120	60
121-200	80
201-500	100
501-1000	125
1,001-Over	150

Table 3 (.03 Standard Error - True Error Rate of .1)

Group Size	Sample Size
1-60	all
61-80	40
81-120	50
121-200	60
201-500	80
501-1000	90
1,001-Over	100

Table 4 Stage I - Date/Time Segments

Example for Feburary

(Selection Interval = 8)

Time	2nd	6th	10th	14th	18th	22nd	26th
2400/0014			X				
0015/0029				X			
0030/0044					X		
0045/0059						X	
0100/0114							X
0115/0129							
0130/0144	X						
0145/0159		X					
0200/0214			X				
0215/0229				X			
0230/0244					X		
0245/0259						X	
0300/0314							X
0315/0329							
0330/0344	X						
0345/0359		X					
0400/0414			X				
0415/0429				X			
0430/0444					X		
0445/0459						X	
0500/0514							X
0515/0529							
0530/0544	X						
0545/0559		X					
0600/0614			X				
0615/0629				X			
0630/0644					X		
0645/0659						X	
0700/0714							X
0715/0729							
0730/0744	X						
0745/0759		X					
0800/0814			X				
0815/0829				X			
0830/0844					X		
0845/0859						X	
0900/0914							X
0915/0929							
0930/0944	X						
0945/0959		X					
1000/1014			X				
1015/1029				X			
1030/1044					X		
1045/1059						X	
1100/1114							X
1115/1129							
1130/1144	X						
1145/1159		X					

Time	2nd	6th	10th	14th	18th	22nd	26th
1200/1214			X				
1215/1229				X			
1230/1244					X		
1245/1259						X	
1300/1314							X
1315/1329							
1330/1344	X						
1345/1359		X					
1400/1414			X				
1415/1429				X			
1430/1444					X		
1445/1459						X	
1500/1514							X
1515/1529							
1530/1544	X						
1545/1559		X					
1600/1614			X				
1615/1629				X			
1630/1644					X		
1645/1659						X	
1700/1714							X
1715/1729							
1730/1744	X						
1745/1759		X					
1800/1814			X				
1815/1829				X			
1830/1844					X		
1845/1859						X	
1900/1914							X
1915/1929							
1930/1944	X						
1945/1959		X					
2000/2014			X				
2015/2029				X			
2030/2044					X		
2045/2059						X	
2100/2114							X
2115/2129							
2130/2144	X						
2145/2159		X					
2200/2214			X				
2215/2229				X			
2230/2244					X		
2245/2259						X	
2300/2314							X
2315/2329							
2330/2344	X						
2345/2359		X					

COUNTING THE UNCOUNTABLE ILLEGALS: SOME INITIAL STATISTICAL SPECULATIONS EMPLOYING CAPTURE-RECAPTURE TECHNIQUES

Clarise Lancaster, Office of the Assistant Secretary for Planning and Evaluation, Department of Health, Education, and Welfare
Frederick J. Scheuren, Social Security Administration

This paper provides some initial statistical speculations on the number of illegal aliens residing in the United States. Our results come from the 1973 CPS-IRS-SSA Exact Match Study [1] which has been conducted jointly by the Census Bureau and the Social Security Administration, assisted by the Internal Revenue Service. Direct estimates are presented only for the age group 18 to 44 years old as of April 1973; however, there is some discussion of ways, using other sources, that one can extend these figures to all age groups and project them forward in time.

Organizationally, the paper is divided into five sections. Section 1 provides a brief introduction to what is known about the nature and magnitude of the illegal alien population. The approach we will take in obtaining estimates for 1973 is described in section 2. Some limitations on the data being used are set forth in section 3. Section 4 discusses the results of the exploratory analyses we have carried out so far. A few conclusions and possible implications for future study are given in section 5.

1. INTRODUCTION

Most of what we know about illegal aliens comes from data on apprehensions (about 800,000 in 1975) which suggest that Mexico is a major source of such individuals.1/ United States and Mexican authorities, however, have,on numerous occasions, cited the unreliability of the apprehension information as indicative of the nature of the total illegal alien population in the U.S. In particular, it is misleading to characterize the illegal alien population in the United States as predominantly male and Mexican based on these apprehension statistics: first, because we are dealing with those who are, in fact, caught, and there is no reason to believe that they are representative of those who are not caught; and, secondly, because Mexican illegal immigration may be substantially different from that of other source countries, mainly Jamaica, the Dominican Republic, Haiti, Korea, the Phillippines, Thailand, and China. It is suspected that both Mexicans and males are over-represented in apprehension data.

Not only is the composition of the illegal alien population unclear from official statistics, but the total number of illegals who are not apprehended is, of course, unknown and is a source of considerable speculation. To see how widely divergent some of the guesses are, it might be worth quoting from a recent article by Hobart Rowen [4] in the Washington Post--

> There are four million illegal aliens in the United States.
>
> There are eight million illegal aliens in the United States.
>
> There are twelve million illegal aliens in the United States.
>
> These are the estimates of [government] officials trying to evolve a policy to deal with illegal immigration. You can pick any one of them, or insert your own number and you will be--they confess--as accurate as they are."The truth is [an official says] that no one knows how many 'illegals' are in the country."

As will be seen later in this paper, our own preliminary investigations suggest that it is the smallest of these figures which is more nearly correct.

2. METHODOLOGY

2.1 General.--The approach we will use to estimate the number of illegal aliens makes use of two sources of information:

1. a sample of the total resident civilian noninstitutional population, including illegal aliens (who were not, however, identifiable as such); and

2. an independent estimate or "count" of the number of persons in the resident civilian noninstitutional population, excluding illegal aliens.

From the sample data, the Capture-Recapture procedure is used to estimate the total resident civilian noninstitutional population including illegal aliens. The independent population total, excluding illegal aliens, is then subtracted from this sample estimate to derive counts for "illegals."

The sample we are using to make estimates is the Census Bureau's March 1973 Current Population Survey (CPS). The capture-recapture technique can be applied to this sample because it has been matched to Internal Revenue Service (IRS) individual income tax records, and Social Security Administration (SSA) earnings and benefit data.

The independent population estimates on which we rely also come from the Census Bureau. They were obtained by adjusting the 1970 Census count for underenumeration and carrying forward the population totals taking account of subsequent aging of the population, births, deaths, and net legal migration [5, 6]. Also excluded from the population estimates were members of the Armed Forces in April 1973 and persons living in institutions [7].

2.2 Capture-Recapture techniques.--In order to explain how we employed the capture-recapture technique, let us examine table 1, which illustrates our approach for the total 18 to 44 year age group. Two observations should be made initially:

1. All the individual cell estimates, except for the lower right-hand corner total, were taken from a random half-sample selected from the 1973 CPS-IRS-SSA Exact Match Study. These were the data with which we started our exploratory analyses. 2/

2. The right-hand corner entry (shown in parenthesis) was obtained by subtracting the remaining cells from the April 1, 1973, Census Bureau estimate (73,893,000) for the total civilian noninstitutional population 18 to 44 (which excludes illegal aliens).

Now the capture-recapture [8], or multiple systems [9], estimation procedure that we used, essentially resolved itself into treating the cell entry in the parenthesis as missing and estimating it from the remainder of the table. Once this was done, the difference between the new entry for the "missing" cell and the original (parenthesized) entry provided our count of "illegals." 3/

To compute the capture-recapture estimate for the missing cell, we employed expression (6.4-15) from [8], that is:

$$m_{222} = \frac{m_{111}\, m_{221}\, m_{122}\, m_{212}}{m_{121}\, m_{211}\, m_{112}},$$

where the cell counts or entries $\{m_{ijk}\}$ are defined by letting i = 1 or 2, depending on whether there is a yes or no, respectively, on the IRS dimension (i.e., whether a person was in a unit with a taxfiler, "yes", or not, "no"); j=1 or 2, depending on whether there is a yes or no on the SSA covered employment dimension; and, finally, k=1 or 2, depending on whether there is a yes or no on the SSA beneficiary dimension.

The above formula for the missing entry m_{222} cannot be interpreted without making a number of (strong) assumptions. Two might be mentioned here:

1. To explain all the interrelationships which exist between the three "captures" (administrative systems), it is enough to look at just the pairwise associations between them. (More technically, the assumption is being made that there is no second-order interaction.)

2. The very same set of "capture" probabilities applies to each individual in the population. Such an assumption would only be tenable if the group we are dealing with were divided into very homogeneous subgroups--something we will discuss in section 4.

2.3 Definition of classifiers.--Some definitions are needed of exactly what we mean by the classifiers in table 1. These are provided in the following paragraphs:

1. SSA beneficiaries.--To be considered an SSA beneficiary, a person had to be receiving benefits in December 1972 (i.e., be in Current Pay Status for that month).

2. SSA covered employment.--To be considered as a covered worker, an individual had to have had taxable SSA wages or self-employment reported for calendar year 1972.

3. Federal income taxfiler.--To be considered a taxfiler, an individual had to have filed a tax return for 1972 on which he was designated as the primary taxpayer. 4/

4. STATS unit.--This is a nuclear family concept used at Social Security to designate individuals in CPS households who would generally be considered interdependent under social insurance programs [10]. The designation, STATS units, stands for "Simulated Tax and Transfer System" units. These units can consist of a single adult 22 years or older, an adult with children under 14, and married couples with or without children. Young adults (14 to 21 years old), depending on their living arrangements, are treated as separate units or as part of a unit containing their parent(s).

TABLE 1.--U.S. civilian noninstitutional population 18 to 44 years old as estimated from the 1973 Census-Social Security Exact Match Study and Census Bureau sources

(In thousands)

In STATS units with persons in SSA covered employment	Total	In STATS units with persons filing Federal income tax returns	
		Yes	No
Overall total....	76,893	67,289	9,604
IN STATS UNITS WITH SSA BENEFICIARIES			
Yes.....	1,321	1,142	179
No.....	509	79	430
NOT IN STATS UNITS WITH SSA BENEFICIARIES			
Yes.....	68,412	63,447	4,965
No.....	6,651	2,621	(4,030)

Note: For definitions of terms used, see section 2.3.

In table 1 above and in the tables used in our subsequent analyses, we do not classify an individual by whether or not he or she was "captured" by one of the administrative systems, but, rather, by whether or not anyone in his or her STATS unit had been so captured. Two (natural) questions arise in this connection: "Why didn't we classify individuals by their own characteristics?" and "How sensitive would our results be if we had done so?"

We didn't classify people just on the basis of their own characteristics for two reasons. First, the STATS unit, by construction, is conceptually more attractive as a classifier of an individual's relationship with regard to the beneficiary and tax systems. Second, by using the STATS unit as a classifier, we expected to increase the overlap

among all three systems, which, in turn, would reduce the probability of having zero cells and, perhaps,make more tenable our assumption of no second-order interaction.

When this paper was delivered in Chicago, we had not yet obtained an answer to the question of how sensitive our results would be if we did the analysis on a person, rather than a STATS unit, basis. The work we have done since then suggests that the results would be very sensitive indeed. The person-based estimates do not actually contradict the STATS unit ones, however. What seems to be happening is that the sampling error of the estimate of the missing cell has increased enormously, principally because much more of the sample was not "captured" by any system.

3. DATA LIMITATIONS

The assumptions which the method requires necessarily impose limitations on our estimates. In addition to these, however, there is also a second set of limitations which arises from the nature of the data on which we are using the method:

1. Survey and matching problems.--The starting point of the administrative record matches was the CPS and not the systems themselves. Problems of non-matches, mismatches, coverage, and non-interview nonresponse must necessarily be considered. (See [11], for example.) It is enough to say here that we believe that these data problems definitely raise interpretive issues, even though major efforts were made to adjust or "correct" for any impacts they might have had [7].

2. Administrative data problems.--The nature of the administrative systems we are using is such that illegal aliens might be less well-represented than their (other) socio-economic characteristics (income level, age, race, sex, etc.) might otherwise suggest. We do not know how serious this is, but it is a problem which we believe would (in the absence of other problems) lead to an underestimation of the total illegal population.

3. Independent population totals.--The Census Bureau population estimates needed for deriving "illegals" are themselves subject to error. Evidence from [12], for example, suggests that there may be a serious understatement in the allowance made for outmigration. For the 18 to 34 year olds this is likely to be the only important error. For the remainder of the 18 to 44 year age group, that is, persons 35 to 44, the undercount totals (Siegel's Preferred Series D) for 1970 are based on a combination of demographic techniques [5, p.6] and not, principally, on vital records, as is true of the younger ages (suggesting that there might be proportionately more error in the older age group).

4. EXPLORATORY ANALYSIS

When this paper was given at the meetings, we were still in the exploratory analysis phase of our research on illegals. In order to be able (at a later date) to do at least some confirmatory analysis, we restricted our attention to half the sample cases in the 1973 Exact Match Study.

4.1 Initial results.--To make more tenable the assumption that the capture probabilities were equal for every individual, we subdivided the age group 18 to 44 into four race-sex subgroups: white males, white females, males of other races, and females of other races. This also has the advantage, as Chandra Sekar and Deming have suggested [13], of tending to lower the overall variance.

Table 1 was repeated for each subgroup separately. The combined tabulation, consisting of 32 cells (four of which were to be treated as missing), was then subjected to "standard" log linear contingency table fitting procedures.5/ Our goal was, of course, the usual one: eliminating those parameters which the analysis showed were unnecessary. In other words, to create a model with fewer parameters which fits well enough to withstand statistical inspection while, at the same time, is sufficiently parsimonious to yield "sturdy" estimates.

Many models were considered before we settled on one to illustrate our results. The model chosen was fit by iterative proportional scaling to the following five sets of marginal totals:

1. Sex
2. Race and taxfiler status
3. Taxfiler status and covered worker status
4. Taxfiler status and beneficiary status
5. Covered worker status and beneficiary status.

Table 2.--Initial Exploratory Model Estimates for April 1973, of Total U.S. Civilian Noninstitutional Population 18 to 44 Years Old by Race and Sex

(Numbers in thousands)

Race and Sex	Total excluding illegals*	Total including illegals	Difference (illegals) Number	Difference (illegals) Percent
Total.....	76,893	79,951	3,058	100.0
Male.....	37,490	39,705	2,215	72.4
Female.....	39,403	40,246	843	27.6
White, total.....	66,673	68,603	1,930	63.1
Male.....	32,689	34,069	1,380	45.1
Female.....	33,984	34,534	550	18.0
Other races, total	10,220	11,348	1,128	36.9
Male.....	4,801	5,635	834	27.3
Female.....	5,419	5,712	293	9.6

(*)Population totals not adjusted for understatement of 1960-73 outmigration.

Once we had obtained our fitted model, we then used the estimates it provided in each of the four race-sex subtables to obtain new entries for the "missing" cells. From the "before" and "after" totals for each race-sex group we then constructed table 2.

4.2 Further results.--We brought a computer terminal with us to the meetings and invited anyone interested in the results in table 2 to try his own hand at still other models. Our basic data set had literally hundreds of dimensions we had not yet looked at. Two we thought most promising were age and income; and we had come prepared to fit models involving these variables if anyone suggested them. As luck would have it, the interactive APL computer service we use was down most of the day of the meeting, and no one was able to take us up on our offer. Matters did not rest at this point, however.

A number of discussions have been held, since the paper was delivered, with various individuals interested in and knowledgeable about illegal alien immigration. From these conversations, we concluded three things. First, we had to provide at least one model which split up the rather broad age group 18 to 44. Second, we had to adjust our initial estimates for the rather serious understatement (over 500,000) in the outmigration estimates used to obtain population totals that excluded illegal aliens. Third, since our initial and improved results had a certain amount of plausibility, they were likely to be believed and used. Therefore, as "responsible" researchers, we had to provide at least some rough idea about the magnitude of the uncertainty surrounding our figures.

In accord with these excellent suggestions, we returned to our exploratory work with the same half sample that was used to obtain table 2. This time we added age as a dimension (18 to 34 and 35 to 44) and looked at models for the 6-way table involving sex, race, age, and the three administrative systems. The model we finally settled on was obtained by fitting the following marginal totals:

1. Sex
2. Race and taxfiler status
3. Age and taxfiler status
4. Taxfiler status and covered employment status
5. Taxfiler status and beneficiary status
6. Covered employment status and beneficiary status.

To test this model, we fit it on the second half of our sample. While the fit (as expected) was not nearly as good on the second half, it still could be accepted at the $\alpha = .05$ level of significance.

Our next step was to combine the two half samples and refit the model on all the data. The estimates obtained in this way are shown in table 3, column (2). The final step we took was to revise the population estimates not including illegals (column (1) of table 3) to account for the understatement of outmigration. The Warren-Peck paper [12], set B estimates were our basic source. These were aged to 1973, the effect of additional outmigrant underestimation between 1970 and 1973 was imputed, and a rough adjustment was made to take account of changes in the foreign student population not originally reflected in [12].6/ The result of these steps is shown below.

Age Group	Understatement of Outmigrants (in thousands)		
	Total	Male	Female
Total.....	568	244	324
18 to 34 years..	440	180	260
35 to 44 years..	128	64	64

Since virtually all of the outmigrants involved were believed to be white, we made the entire adjustment in that racial group.

Table 3.--Overall Revised Model Estimates for April 1973 of Total U.S. Civilian Noninstitutional Population 18 to 44 Years Old by Race and Sex

(Numbers in thousands)

Race and Sex	18 to 34 Years of Age			35 to 44 Years of Age		
	Total excluding illegals*	Total including illegals	Difference (illegals)**	Total excluding illegals*	Total including illegals	Difference (illegals)**
Total.....	53,401	56,583	3,182	22,924	23,627	703
Male.....	25,973	27,974	2,001	11,273	11,681	408
Female.....	27,428	28,609	1,181	11,651	11,946	295
White, total.....	46,198	48,379	2,181	19,910	20,304	394
Male.....	22,613	23,918	1,305	9,834	10,038	204
Female.....	23,585	24,461	876	10,076	10,266	190
Other races, total.	7,203	8,204	1,001	3,014	3,323	309
Male.....	3,360	4,056	696	1,439	1,643	204
Female.....	3,843	4,148	305	1,575	1,680	105

(*)Adjusted for outmigration as explained in the text.
(**)This estimate differs from that in table 2 due to the adjustment for outmigrants discussed in the text, to the fact that the whole sample is being used, not just half, and to the fact that the models fit in the two cases are different.

4.3 Crude measures of uncertainty.--It is a formidable, perhaps impossible, task to do a "good" job of assigning measures of uncertainty to the entries for "illegals" in table 3. We have to obtain the approximate sampling errors of the estimates, quantify the impact of the nonsampling errors, and assess the robustness of the figures to possible failures in the assumptions underlying our application of the capture-recapture method.

Time considerations precluded our making more than a crude attempt to quantify the uncertainty surrounding the estimates in table 3. Perhaps we should not even have tried, since subjective judgments play such an important role in our assessments and, undoubtedly, other researchers may reach quite different conclusions.

Table 4 provides the rough confidence bounds we constructed.7/ Notice that they are not symmetric, reflecting our belief that the counts of "illegals" in table 3 may be downwardly biased. The bounds also are quite far apart. This is in keeping with the early stage at which our analysis stands. Further research probably would lead to estimates with narrower bounds of uncertainty.

Table 4.--Subjective 68 percent Confidence Intervals for the Overall Revised Model Estimate of the Number of Illegal Aliens 18 to 44 Years of Age in April 1973 by Age, Race and Sex

(In thousands)

Race and Sex	18 to 44 Years of Age		18 to 34 Years of Age		35 to 44 Years of Age	
	Lower	Upper	Lower	Upper	Lower	Upper
Total.....	2,904	5,722	2,438	4,574	466	1,148
Male.....	2,046	3,318	1,726	2,689	320	629
Female.....	858	2,404	712	1,885	146	519
White, total.....	1,961	3,724	1,715	3,052	246	672
Male.....	1,282	2,077	1,133	1,735	149	342
Female.....	679	1,647	582	1,317	97	330
Other races,total..	943	1,998	723	1,522	220	476
Male.....	764	1,241	593	954	171	287
Female.....	179	757	130	568	49	189

5. SOME CONCLUSIONS AND IMPLICATIONS

According to the overall model shown in table 3, there were some 3.9 million resident "illegals" 18 to 44 years of age in April 1973. Rough, subjective, 68 percent confidence bounds on this estimate (from table 4) suggest that the actual value could be anything from 2.9 million to 5.7 million. Generally speaking, such widely (wildly?) varying speculations would cause most people to make no further demands on the present results. We certainly would not wish to do so were it not for the fact that the questions of most interest are--

> "How many illegals were there, altogether, in 1973?"
> "How much has the total increased since 1973?"

We cannot offer any statistical speculations of our own on these questions, but it might be worth mentioning how others have answered them. First, David North, in [14], cites various studies which ..."suggest that the 18-44 age range would cover most, but not all, of the illegal aliens; a 10% upward adjustment would appear appropriate. ..." On the second question, we turn to some conclusions of Alex Korns [15], who has examined the relationship between the BLS establishment and CPS employment series for nonagricultural wage and salary jobs. He notes that while there may have been a sharp rise in illegal alien employment during the business expansion of 1964-1969, there appears to be no sustained increase since then.

With these two outside sources in mind, we feel reasonably comfortable in restating the assertions about the number of "illegals" that Rowen quoted:

> There are probably not twelve million illegal aliens in the United States.
>
> There are probably not eight million illegal aliens in the United States.
>
> There may, however, be about four million illegal aliens in the United States.

AN AFTERWORD

We debated whether or not to submit this paper to the Proceedings. The subject is, after all, important and controversial; hence, it deserves a careful, studied treatment. Unfortunately, time and resource constraints intervened. Our results, therefore, are quite preliminary and could be misleading if taken too seriously.

Ultimately, what persuaded us to give the paper and, then, have it published was an expectation that other statisticians interested in "illegals" would learn about the 1973 Exact Match Study data base and use it in their own research. The public-use files from the study are now available and may provide the means to do the complete, thorough job that the subject deserves. We would be more than happy to assist in any such effort.

ACKNOWLEDGEMENTS AND FOOTNOTES

The authors would like to thank several individuals for sharing their expertise on illegal aliens: David North, Alex Korns, Muffie Houstoun, and especially Robert Warren. We also benefitted considerably from discussions with Jeff Passel and Jacob Siegel at the Census Bureau after the paper was delivered at the Chicago meetings. Editorial and other assistance was provided by Ben Bridges, H. Lock Oh, Linda DelBene and, especially, Wendy Alvey. The typing was done by Joan Reynolds and Helen Kearney.

We would also like to take this opportunity to mention two points about the title of our paper. First, "Counting the Uncountables" is apparently an irresistable phrase. The Illegal Alien Study Design report [3], for example, uses the expression, something we were not aware of when we chose it ourselves. The Design report also suggests that the well-known "Capture-Recapture" technique be employed to estimate the number of illegal aliens. In doing so, the authors of that report add a graceful apology, with which we concur, for the necessity of using such (customary) terminology with respect to this population.

1/ The authors have relied primarily on [2] and [3] for the brief overview of the illegal alien immigration situation in this section.

2/ The estimates were obtained by using twice the "Final" administratively weighted [7] sample figures from rotation panels entering the survey in March for the first, third, sixth or eighth time.

3/ In the more general settings later in section 4, the "count of illegals" is obtained by calculating the difference between the model estimated total population derived from the sample (which includes "illegals") and the Census supplied population (where "illegals" are excluded). It might be mentioned also that just because we sometimes calculate our estimates from the "missing" cell does not imply that this is where all the illegals will be found. Quite the contrary. If none of the "illegals"

were ever "captured" by the administrative systems, then our procedure simply would not work.

4/ For nonjoint returns, there was considered to be only one taxpayer; for joint returns filed by married couples, there were two. In such cases, the husband was designated as the primary taxpayer.

5/ Actually, standard log linear procedures require simple random sampling. The CPS sample design and estimation procedures were such that we had to modify the ordinary minimum discrimination information (maximum likelihood χ^2) test statistic by dividing by the product of the base weight for the half sample (3,200) times a preliminary estimate of the design effect (taken to be quite large, about 3). The data for both half samples is available upon request.

6/ The updating and adjustments were prepared with the help of Robert Warren.

7/ The actual steps we went through to obtain these crude bounds are available upon request.

REFERENCES

[1] U.S. Social Security Administration, Studies from Interagency Data Linkages, Reports Nos. 4 and following.

[2] Domestic Council Committee on Illegal Aliens, Preliminary Report, December 1976.

[3] U.S. Immigration and Naturalization Service, Illegal Alien Study Design, Vol. 1-Final Report, May 1975.

[4] Rowen, H., "Illegal Alien Dilemma," Washington Post, Section A, p. 19, July 21, 1977.

[5] Siegel, J., Estimates of Coverage of Population by Sex, Race and Age: Demographic Analysis, 1970 Census of Population and Housing: Evaluation and Research Program, PHE(E)-4, 1974.

[6] U.S. Bureau of the Census, "Population Estimates and Projections," Current Population Reports, Series P-25, No. 614.

[7] Scheuren, F., "Methods of Estimation for the 1973 Exact Match Study" (unpublished working paper to appear in the series Studies from Interagency Data Linkages).

[8] Bishop, Y., Fienberg, S., and Holland, P., Discrete Multivariate Analysis: Theory and Practice, Cambridge: MIT Press, 1975, Chapter 6.

[9] Marks, E., Seltzer, W., and Krotki, K., Population Growth Estimation: A Handbook of Vital Statistics Measurement. New York: The Population Council, 1974.

[10] Projector, D., Millea, M., and Dymond, K., "Projection of March Current Population Survey: Population Earnings, and Property Income, March 1972 to March 1976," Studies in Income Distribution, Report No. 1, Social Security Administration, 1975.

[11] Yuskavage, R., Hirschberg, D., and Scheuren, F., "The Impact on Personal and Family Income of Adjusting the Current Population Survey for Undercoverage," 1977 American Statistical Association Proceedings, Social Statistics Section.

[12] Warren, R. and Peck, J., "Emigration from the United States: 1960 to 1970," paper presented at the Population Association's annual meeting in Seattle, Washington, July 17-19, 1975.

[13] Chandra Sekar, C. and Deming, W., "One Method of Estimating Birth and Death Rates and the Extent of Registration," Journal American Statistical Association, Vol. 44, 1949, pp. 101-115.

[14] North, D., "Manpower Policy and Immigration Policy in the United States: An Analysis of a Nonrelationship," Chapter IV and Appendix D. (This report is to be published shortly by the National Commission for Manpower Policy.)

[15] Korns, A., "Coverage Issues Raised by Comparisons between CPS and Establishment Employment," 1977 American Statistical Association Proceedings, Social Statistics Section.

QUEENING IN MASTER CHESS TOURNAMENTS: 1867-1970

Ernest Rubin, American University and the University of the District of Columbia

Introductory Remarks

According to the Oxford Universal Dictionary, "queening" in chess was in force as early as 1440.1/ Currently, chess rules require that a pawn be promoted, when it has reached the eighth rank, to one of four pieces, i.e., a queen, a rook, a bishop or a knight.2/ For over a century this manner of queening has been adopted in master chess play.3/

The purpose of this essay is to describe tentative answers to certain questions regarding queening in master chess tournaments. Fourteen chess competitions, covering the period 1867 to 1970, were studied. Over 1,200 games were examined in detail and 89 queenings were identified in 70 games.4/ The following results are based on this sample.

Frequency of Queening

Table 1 provides information on each tournament, it indicates total games played and the games in which one or more queenings occurred. Somewhat less than 6 percent of total games involve queening. The lowest percentage, 1.5 percent, occurred in the U. S. Championship Tournament of 1969; the highest percentage, 10 percent, took place in two tournaments, that of Hastings, 1922 and New York, 1924.

The data, viewed as a time series, suggest a decline, since 1935, in the relative frequency of queenings in master chess tournaments. Many more tournaments, however, must be studied before definitive trends of queening in master chess can be established.5/

Outcome After Queening

What is the relation between queening and the outcome of a game? The answer to this question is shown in Table 2. For the sample of 70 games, draws occurred 13 times or 18.6 percent while white won 29 times or 41.4 percent and black 28 times or 40 percent.6/

It will be noted that of the 14 tournaments studied, the Vienna 1903 and Baden 1914 competitions were gambit tournaments. The conditions specified in the Vienna 1903 tournament required white, the first player, to open with the King's Gambit (i.e., P-K4 and then P-KB4); and black, the second player, to accept the King's Bishop Pawn, by responding, respectively, to white by P-K4 and PxP.7/ The Baden 1914 tournament conditions were less stringent than those of Vienna 1903. Only Gambit openings were allowed, but the acceptance of the Gambit was not mandatory; a player had to obtain permission from the tournament committee, specifying the variation he would play in declining the Gambit.8/

Separating the sample of 70 queening games by type of tournament origin, there were 58 queening games in 12 regular tournaments and 12 such games in 2 Gambit tournaments. In the "regular" tournaments white won 27 games, black 21 games and 10 games were drawn; the corresponding percentages are white .47, black .36 and drawn .17. In Table 3, comparable data are provided for total games played in the 14 tournaments. For the total of 1,049 regular tournament games the white won 35 percent, black 30 percent and 35 percent were drawn. In the two gambit tournaments of 179 games, the corresponding percentages are white 31 percent, black 36 percent and 33 percent drawn. The data for the Gambit tournaments suggest a substantial advantage for Black, both as to queening and as to winning.

Methods of Queening and Nature of Pawn Promotion

Two ways determine pawn promotion. A pawn may reach a square on the eighth rank by advancement or by capture of an enemy piece occupying that equal.

Of the 89 queenings that occurred in the 14 tournaments under investigation, 83 (or about 93 percent), were the result of pawn advancement. Only six queenings came about when an enemy piece was captured. (Table 4). In the 12 regular tournaments, white pawns queened 42 times and black pawns 34 times. Queening after capture occurred twice for white and for black. For the two gambit tournaments, there were only two queenings by white and eleven by black. In this group only two queenings, by black, occurred following capture.

Since the queen is the most valuable piece, it is no surprise that pawn promotion to queen is the overwhelming choice. For the 89 pawn promotions that occurred in 70 games, 87 were to queens and only 2 pawn promotions were to knights. In this sample not a single pawn was promoted to rook or bishop.

Distribution of Queening Events and Their Outcomes

In theory, eight pawns on each side have the potential of becoming queens. I know of no game ever played in which one side had nine queens and I would conjecture this result to be impossible. Chernev refers to a game with five games, 3 white and 2 black, and to a world championship game with four queens, 2 white and 2 black.9/

Table 5 provides the distribution of queening events in 70 games, with a breakdown for color and type of tournament. In 55 of these games, only one side queened, i.e., white 28 and black 27. If we consider regular tournaments, there were 58 games with queening events in which only white queened, 27 times, and only black queened, 19 times. For the two gambit tournaments in which there were 12

queening events, only black queened, 9 times, only white, 2 times; in one game each side queened once. Thus, the number of queenings in master play is predominantly of the singular type. There were only four games in which three pawns were queened.

Of interest is the relationship between queening, color, game outcome and type of tournament. This information is provided in Table 6. When only white queened once or more, in 29 games, white won 23, lost 1 and drew 5. The comparable information for black in 28 games is 22 wins, 4 losses and 2 draws, indicating a slight advantage to white. In 9 games, white and black queened once, with the results that 5 games were drawn, white won once and black three times.

Somewhat modified results obtain by considering the type of tournament. In Table 6 consider the 58 results for the twelve regular tournaments. White performance appears appreciably better in regular tournament competition with regard to queening. In 27 games only white queened once or more, resulting for white in 22 wins, 1 loss and 4 draws (a score of .889). When only black queened once or more in 19 games, black won 15, lost 3 and drew 1 (a score of .816).

Queening Pawns and Queening Squares

Which pawns queen most frequently in tournament competition? Table 7 provides the data based on 89 queenings in the 70 game sample. For white and for black, the queen rook pawns queened most frequently, followed by the king rook pawn. Together the queen and king rook pawns accounted for 31 queenings out of a total of 89 queenings or better than 34 percent. Second in importance are the knight pawns, with 23 queenings or somewhat over 25 percent of total queenings. Tied for lowest were the king and king bishop.10/ Pawns with 16 queenings or around 18 percent.

These findings are consistent with chess practice and theory. On a priori grounds the expectation is that the center pawns would queen relatively infrequently because these pawns are most often captured or exchanged.11/ Pawns on the flanks, i.e., the rook pawns, however, are usually the survivors of openings and of early middle games. On the average, queenings occurred on the 51st move for white and the 46th move for black.12/

On which squares do queenings occur? The results (Table 7), for the 89 queenings were somewhat unexpected in view of the distribution of the queening pawns. On the high side were the queen knight and king knight squares. These squares accounted for 28 queening sites, followed closely by the corresponding rook squares. The king and queen squares were of minimal significance, constituting the site of only 12 queenings.

The correspondence between the original position of the queening pawns and their queening squares, using rank correlation, was 80 percent. The knight, as well as bishop, queen, and king, squares provide three ways of pawn access (by capture of an enemy piece from two adjacent columns or by advancement); the rook squares permit only two ways of pawn access. It is possible for a king rook pawn to queen on the queen knight square, in effect, a shift of seven columns. In practice, however, the moves of a queening pawn are at most two columns from the original column position; and most queening pawns remain in the original column.

Queening and Castling

The data provided in Table 8 relate queening and castling in the 70 game sample. In 29 games only white queened and in 28 games only black queened; in 13 games white and black queened. Of the 57 games in which queening was accomplished by either white or black, castling took place 46 times. Thus, in 11 games the queening player did not castle (5 white and 6 black). For the 13 games in which both sides queened, the castling move was completed 26 times, i.e., 13 times by white and 13 times by black.

As indicated in Table 8, castling on the king's side predominated, 62 times, compared with 10 times on the queen's side and 11 times in which the castling did not occur.

The distribution of queening by side of board is much more balanced than the distribution of castling. Queening and castling on the king side occurred 30 times and on the queen side 32 times. When castling occurred on the queen side, queening was equally divided, 5 on the king and 5 on the queen side. In the 11 games in which castling did not occur, 4 queenings took place on the king side and 7 queenings occurred on the queen side.

While the preponderance of castling on the king side was to be expected, it seems rather unusual that queening should occur almost equally on both sides of the board. Further samples will have to be examined, however, before any conclusions of the relationship between queening and castling sides can be asserted with any statistical reliability. The results obtained in this paper are not based on a random sample and a sample of 70 "queening" games is not large enough for the number of variables under discussion. These findings are, perhaps, suggestive or indicative of what may possibly be expected from a more extensive sample.

Summary of Principal Findings

The principal results of this analysis suggest that queenings in master chess between 1867 and 1970 occurred in about 5½ percent of the games played. There is an indication, however, that in master tournaments since 1935 the percentage of queenings in tournaments has fallen below 4 percent.

In general, the side to queen first won but it should be noted that 13 of 70 "queening" games resulted in draws, and in two games the queening side lost. Of interest is the result that under-

promotion, to knight, was made twice while the remaining 87 promotions were to queens. For both the white and black sides, the queen rook pawn queened most frequently, followed by the king rook pawn and the king and queen knight pawns.

This study should be regarded as a preliminary investigation into the phenomenon of queening in master chess tournaments. Of interest would be a comparable examination of queening in world chess championship matches. A considerably larger sample would enable us to relate chess openings to queening as well as to establish statistically defensible conclusions.

In terms of queening, it becomes very obvious that there is a variation in the value of the pawns and perhaps further study should be made of pawn evaluation in terms of their initial position.10/

Finally, we must reiterate that the results of this study are subject to the strictions that (a) the sample is not random, and (b) the sample is small in terms of the number of considered variables. It is hoped, however, that this study will be a spur to further work on this problem.

Sources:

a. Dundee Centenary Tournament, 1967 and British Chess Association Congress, Dundee 1867 by R. G. Wade (The Chess Player, Nottingham, England, 1967)

b. The Hastings Chess Tournament, 1895, edited by Horace F. Cheshire, (Dover Publications, N. Y., 1962)

c. Vienna Gambit 1903, annotations by George Marco (The Chess Player, Omaha, Nebraska, 1967)

d. The International Chess Congress, St. Petersburg, 1909, by Dr. Emanuel Lasker, (Dover Publications, N. Y., 1971)

e. Baden 1914 Chess Gambit Tournament, (James R. Schroeder, Cleveland, Ohio. 1972)

f. Hastings International Masters' Chess Tournament 1922, ed. by W. H. Watts (Dover Publications, N. Y., 1968)

g. London International Chess Congress 1922, ed. by W. H. Watts (Dover Publications, N. Y., 1968)

h. New York International Chess Tournament 1924, ed. by Hermann Helms (Dover Publications, N. Y., 1961)

i. Nottingham International Chess Tournament, 1936, ed. by W. H. Watts (Dover Publications, N. Y., 1962)

j. Soviet Chess Championship 1941, by M. M. Botwinnik (Dover Publications, N. Y., 1973)

k. Canadian Centennial Grand Masters Chess Tournament 1967, ed. Ken Smith (Chess Digest, Dallas, Texas, 1968

l. 1969 United States Chess Championship and World Championship Zonal Qualifier, by Morton Siegel (United States Chess Federation, Newburgh, N. Y., 1970

m. The Match of the Century: U.S.S.R. v. Rest of the World, by David N. L. Levy (The Chess Player, Nottingham, England, 1970)

Notes and References

1. "Queen...8. In games a. Chess...the positions on the board attained by a pawn, when it is queened .1440," The Oxford Universal Dictionary on Historical Principles, 3rd ed., revised Clarendon Press, Oxford, 1955), p. 1638.

2. Harkness, Kenneth, Official Chess Rulebook, (David McKay, N. Y., 1970): "On reaching the last rank a pawn must be immediately exchanged, as part of the same move, for a queen, a rook, a bishop, or a knight of the same color as the pawn, at the player's choice...", p. 26.

3. The precise evolution of the queening rule is difficult to establish. For a brief discussion of this point see "Pawn Promotion" in The Encyclopedia of Chess compiled by Anne Sunnucks (St. Martin's Press, N. Y., 1970), pp. 348-349. Also of interest is Sunnucks' entry "Queen, The" at pp. 394-395. In Volume VIII A New English Dictionary on Historical Principles, ed. by James A. H. Murray (Oxford Clarendon Press, 1914), the following comment appears at page 41 with reference to queening: "1789 Twiss Chess II. 155 Damer le Pion, literally to queen the Pawn, is a French expression. 1797 Encycl. Brit. (ed. 3) IV. 640 notes, To queen is to make a queen." I believe that the current rule on queening was generally established in the first half of the nineteenth century.

4. In this survey of 1,228 games, the score of each game, as given in the sources, was read at least three times. It cannot be claimed, however, that this survey is error-free, in that a few games involving queening may have been missed.

5. Changing styles of chess play between 1867 to 1970, primarily from open to close games, may account for relative changes in the queening phenomenon. Another factor for consideration in this connection is that of the selective basis for player tournament participation. In the last fifty years selection has become successively differentiated because of the improved methods of rating masters, international masters and grandmasters.

6. In Chess Life and Review, May 1976, p. 277, reference is made to an article "White or Black?" by J. Alonso which appeared in Ajedrez Canario (Spain) October 1974. Alonso sampled chess games for the period 1951 to 1970

and found that "...the expectation is 31% for White, 22% for Black, and 47% Drawn."

7. Vienna Gambit 1903 (Spence Tournament Classics, The Chess Player, Omaha, Nebraska, 1967), p. 54.

8. Baden 1914 Chess Gambit Tournament (J. R. Schroeder, Cleveland, Ohio, 1972), p. 4.

9. Irving Chernev, Wonders and Curiosities of Chess (Dover Publications, Inc., N. Y. 1974), pp. 127-128 and p. 148.

10. Encyclopedia Britannica (1st Edition Bell and MacFarquhar, Edinburgh, 1771), entry "Chess" "...The difference of the worth of pawns is not as great as that of noblemen; only, it must be observed, that the king's bishop's pawn is the best in the field...," Vol. 2, p. 182.

11. Ernest Rubin, "Life and Death of a Chess Piece," The American Statistician, April 1963, pp. 20-21.

12. In the games that only white queened, the average length of the game was 61 moves compared with the average length when only black queened of 51 moves.

Table 1. Queening in Selected Master Chess Tournaments: 1867-1970

Date	Competition	Total Games Played	Games with one or more queenings	Percent
1867	Brit. Chess. Cong. Dundee[a]	30	1	.033
1895	Hastings[b]	230	13	.057
1903	Vienna Gambit[c]	89	8	.090
1909	St. Petersburg, International [d]	175	11	.063
1914	Baden Gambite[e]	90	4	.044
1922	Hastings, International[f]	30	3	.100
1922	London, International [g]	120	9	.075
1924	New York, International[h]	110	11	.100
1936	Nottingham, International[i]	105	2	.019
1941	Soviet Chess Championship[j]	60	3	.050
1967	Centenary Tourn. Dundee[a]	38	1	.028
1967	Canadian Centennial[k]	45	2	.044
1969	U.S. Championship[l]	66	1	.015
1970	U.S.S.R. v Rest of World[m]	40	1	.025
Total (14 tournaments)		1,228	70	.057
Subtotal (12 Regular Tournaments)		1,049	58	.055
Subtotal (2 Gambit Tournaments)		179	12	.067

Table 2. Outcome After Queening in 70 Master Chess Games, by Competition: 1867-1970

Date	Competition	Games with 1 or more queening	Won by White	Won by Black	Drawn
1867	Dundee	1	0	1	0
1895	Hastings	13	5	6	2
1903	Vienna Gambit	8	2[a]	4	2
1909	St. Petersburg	11	7	3	1
1914	Baden Gambit	4	0	3	1
1922	Hastings	3	1	1	1
1922	London	9	1[a]	4	4
1924	New York	11	7[a]	2	2
1936	Nottingham	2	2[a]	0	0
1941	Soviet Champion	3	1	2	0
1967	Dundee Cent.	1	1	0	0
1967	Canadian Cent.	2	1[b]	1	0
1969	U.S. Champion	1	1	0	0
1970	U.S.S.R. v World	1	0	1	0
Total	-	70	29	28	13

a/ In one game, only black queened and lost.

b/ In one game, only white queened and lost.

Table 3. Outcomes of 1,228 Games in 14 Selected Masters' Chess Tournaments, by Color: 1867-1970

Date	Competition	Total Games	Games Won by White	Games Won by Black	Draws
1867	Dundee	30	14	15	1
1895	Hastings	230[a]	85	87	58
1903	Vienna Gambit	89[a]	32	37	20
1909	St. Petersburg	175	65	55	55
1914	Baden Gambit	90	24	27	39
1922	Hastings	30	8	10	12
1922	London	120	51	37	32
1924	New York	110	31	41	38
1936	Nottingham	105	39	24	42
1941	Soviet Champion	60	18	13	29
1967	Dundee, Cent.	38	12	12	14
1967	Canadian Cent.	45	8	9	28
1969	U.S. Champion.	66	19	11	36
1970	U.S.S.R. v World	40	13	6	21
Totals, 14 tournaments		1,228	419	384	425
Totals, 12 Regular tournaments		1,049	363	320	366
Totals, 2 Gambit tournaments		179	56	64	59

a/ Does not include 1 game forfeited because of illness.

Table 4. Methods of Queening, by Type of Tournament and by Color, in 70 Master Chess Games: 1867-1970

Method of Queening	Queening		
	White	Black	Total
12 Regular Tournaments			
By Advancing Pawn to 8th Rank	40	32	72
By Pawn Capture on 8th Rank	2	2	4
Sub-total.....	42	34	76
2 Gambit Tournaments			
By Advancing Pawn to 8th Rank	2	9	11
By Pawn Captive on 8th Rank	0	2	2
Sub-total.....	2	11	13
14 Tournaments			
By Advancing Pawn to 8th Rank	42	41	83
By Pawn Capture on 8th Rank	2	4	6
Total.....	44	45	89

Table 5. Queening Events, by Type of Tournament and by Color, in 70 Master Chess Games: 1867-1970

Queening Events (per Game)	Number of Games	Queenings		
		White	Black	Total
12 Regular Tournaments				
1 White, 0 Black.....	26	26	0	26
2 White, 0 Black.....	1	2	0	2
0 White, 1 Black.....	18	0	18	18
0 White, 2 Black.....	1	0	2	2
1 White, 1 Black.....	8	8	8	16
1 White, 2 Black.....	3	3	6	9
2 White, 1 Black.....	1	2	1	3
Sub-Total.....	58	41	35	76
2 Gambit Tournaments				
1 White, 0 Black.....	2	2	0	2
2 White, 0 Black.....	0	0	0	0
0 White, 1 Black.....	9	0	9	9
0 White, 2 Black.....	0	0	0	0
1 White, 1 Black.....	1	1	1	2
1 White, 2 Black.....	0	0	0	0
2 White, 1 Black.....	0	0	0	0
Sub-Total.....	12	3	10	13
14 Tournaments				
1 White, 0 Black.....	28	28	0	28
2 White, 0 Black.....	1	2	0	2
0 White, 1 Black.....	27	0	27	27
0 White, 2 Black.....	1	0	2	2
1 White, 1 Black.....	9	9	9	18
1 White, 2 Black.....	3	3	6	9
2 White, 1 Black.....	1	2	1	3
Total.....	70	44	45	89

Table 6. Outcome in 70 Master Chess Games, by Color, Queening Event, and Type of Tournament: 1867-1970

Queening Event	White Won	White Lost	White Drew	White Total	Black Won	Black Lost	Black Drew	Black Total
12 Regular Tournaments:								
1 White, 0 Black	21	1	4	26	1	21	4	26
2 White, 0 Black	1	0	0	1	0	1	0	1
0 White, 1 Black	3	14	1	18	14	3	1	18
0 White, 2 Black	0	1	0	1	1	0	0	1
1 White, 1 Black	1	3	4	8	3	1	4	8
1 White, 2 Black	0	2	1	3	2	0	1	3
2 White, 1 Black	1	0	0	1	0	1	0	1
Sub-total	27	21	10	58	21	27	10	58
2 Gambit Tournaments:								
1 White, 0 Black	1	0	1	2	0	1	1	2
2 White, 0 Black	0	0	0	0	0	0	0	0
0 White, 1 Black	1	7	1	9	7	1	1	9
0 White, 2 Black	0	0	0	0	0	0	0	0
1 White, 1 Black	0	0	1	1	0	0	1	1
1 White, 2 Black	0	0	0	0	0	0	0	0
2 White, 1 Black	0	0	0	0	0	0	0	0
Sub-total	2	7	3	12	7	2	3	12
All Tournaments:								
1 White, 0 Black	22	1	5	28	1	22	5	28
2 White, 0 Black	1	0	0	1	0	1	0	1
0 White, 1 Black	4	21	2	27	21	4	2	27
0 White, 2 Black	0	1	0	1	1	0	0	1
1 White, 1 Black	1	3	5	9	3	1	5	9
1 White, 2 Black	0	2	1	3	2	0	1	3
2 White, 1 Black	1	0	0	1	0	1	0	1
Total	29	28	13	70	28	29	13	70

Table 7

Table 7. Queening Pawns and Squares in 70 Master Chess Games, by Color and Type of Tournament: 1867-1970

Original Pawn Position	White Tournaments All	White Tournaments Regular	White Tournaments Gambit	Queening Square	Tournaments All	Tournaments Regular	Tournaments Gambit
Queen Rook	9	9	0	Queen Rook 8	8	8	0
Queen Knight	4	4	0	Queen Knight 8	4	4	0
Queen Bishop	5	4	1	Queen Bishop 8	7	5	2
Queen	4	3	1	Queen 8	1	1	0
King	2	1	1	King 8	4	4	0
King Bishop	5	5	0	King Bishop 8	5	5	0
King Knight	8	8	0	King Knight 8	9	9	0
King Rook	7	7	0	King Rook 8	6	6	0
Total White	44	41	3	--	44	42	2
Black							
Queen Rook	9	8	1	Queen Rook 8	9	6	3
Queen Knight	7	4	3	Queen Knight 8	7	6	1
Queen Bishop	5	3	2	Queen Bishop 8	7	5	2
Queen	5	5	0	Queen 8	6	6	0
King	6	5	1	King 8	1	1	0
King Bishop	3	3	0	King Bishop 8	4	4	0
King Knight	4	2	2	King Knight 8	8	5	3
King Rook	6	5	1	King Rook 8	3	2	1
Total Black	45	35	10	--	45	35	10
Total							
Queen Rook	18	17	1	Queen Rook 8	17	14	3
Queen Knight	11	8	3	Queen Knight 8	11	10	1
Queen Bishop	10	7	3	Queen Bishop 8	14	9	5
Queen	9	8	1	Queen 8	7	7	0
King	8	6	2	King 8	5	5	0
King Bishop	8	8	0	King Bishop 8	9	9	0
King Knight	12	10	2	King Knight 8	17	14	3
King Rook	13	12	1	King Rook 8	9	8	1
Total	89	76	13	--	89	76	13

Table 8. Queening and Castling, by Color and by Side of Board, in 70 Master Chess Games: 1867-1970[a/]

Queening Games	Castling Side King	Castling Side Queen	Did Not Castle	Game Total
Only by White				
On King Side	11	2	3	16
On Queen Side	10	1	2	13
Total	21	3	5	29
Only by Black				
On King Side	7	3	1	11
On Queen Side	11	1	5	17
Total	18	4	6	28
By White and by Black				
On King Side	12	0	0	12
On Queen Side	11	3	0	14
Total	23	3	0	26[b/]

a/ In 6 games, queening by the same side occurred twice. The scored queening (2 by white and 4 by black) has been omitted from this table.

b/ The total of 26 queenings represents 13 games because white and black queened in the same game.

FOURIER ANALYSIS OF THE NUMBER OF PUBLIC LAWS 1789 - 1976

David L. Farnsworth, Eisenhower College
Michael G. Stratton, GTE Sylvania

1. Introduction. In an earlier study (reference [1], hereafter referred to as EFF) the number of public laws (NPL) for each United States Congress was analyzed using moving averages, linear regression and similar methods. A centered five Congress moving average of NPL which had been subsequently detrended (called cyclic numbers in EFF) was used as the dependent variable. Three distinct eras of different levels of productivity of laws were perceived and examined separately. Various political variables such as the President's percent of the popular vote and the percent of the Senate which is Republican were treated as independent variables and their degree of correlation with the processed NPL measured.

Here Fourier analysis methodology (see [2]) is used to re-examine the behavior of NPL over time.

There has been some debate concerning whether the number of public laws has meaning [1]. The actual count of NPL is unquestioned since the identity of public laws is clear. Roughly speaking a public law is what one usually thinks of as a law passed by Congress and not vetoed by the President. Other types of actions which require voting by Congress are private laws and internal business such as votes on adjournment. A private law might, for example, allow one particular person to immigrate.

In a sense we are examining the quantity of output of laws alone and trying to make some order out of it. In this study we ignore any independent variables - political, economic, social or others. We are assuming that each law has the same weight in our count or that the total NPL for each Congress is metric data.

The goal herein is to look for periodic components of NPL over time through Fourier decomposition. This is an independent means of validating whether the eras are meaningful or not. This is done without massaging the data or using any preconceived or a priori notions.

We assume that there is stability in the system. That is, regardless of changes such as those in committee structures in Congress, the advent of social legislation, etc., the overall political structure is unchanged over all 94 Congresses. Hence, the output (i.e. NPL) can be inquired into as a single data set.

2. Decomposition into Eras. The present authors agree with EFF in its decomposition of the Congresses into three eras based upon an examination of the five Congress moving average of NPL. The three eras can be extended to include all Congresses with no transition periods between eras. The three eras in EFF are: Era 1 - Congresses 3-34, Era 2 - Congresses 41-66, and Era 3 - Congresses 70-85. For example, the NPL for Congresses 35 and 36 are each within one (Era 1) standard deviation of the mean NPL of Era 1, but the NPL for Congresses 37 through 40 are each more than five (Era 1) standard deviations from that mean. On the other hand, the NPL for Congresses 35 and 36 are both more than 2.8 (Era 2) standard deviations from the mean NPL of Era 2, and all of Congresses 37 through 40 have NPL within about one (Era 2) standard deviation of the mean NPL of Era 2.

The extended Era 1 is composed of Congresses 1 through 36. This adds two Congresses onto each end of the EFF Era 1. These two were deleted in EFF as an artifact of the five Congress moving average. Similarly, the new Era 2 commenses with Congress 37 and ends with 66. Era 3 begins with Congress 67 and ends with the last Congress (94). There is no apparent reason for the deletion of the more modern Congresses 86 through 94 from the third era. In particular the NPL of each of these nine Congresses are within 1.3 (Era 3) standard deviations of the mean NPL of Era 3. However, the downtrend since the 84th Congress should be noted.

3. Autocorrelation in Each Era. The authors are indebted to Ms. P. Paolotto of Colgate University for her comments and initial calculations which form the basis of this section. The eras of EFF are used here.

The von Neumann or Durbin - Watson ratio of lag one measures autocorrelation and also randomness [3]. This statistic in its non-circular form for data $x_1, x_2, \ldots, x_n$ is

$$\frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^{n-1} (x_i - \bar{x})^2}$$

and ranges from zero to four. The value of two represents randomness or no autocorrelation.

The circular autocorrelation coefficient r with lag one enjoys the exact relationship $v = 2(1-r)$ with the circular von Neumann ratio v with lag one. This relationship holds approximately for the non-circular definitions. We utilize the von Neumann ratio instead of the equivalent autocorrelation coefficient.

The von Neumann ratios for the NPL of Eras 1, 2 and 3 are 1.55, 1.42 and 1.75 respectively. For the residuals of the linear least-squares fit of NPL the ratios are 1.93, 1.61 and 1.73. None are significantly different from 2.00 at level 0.10. However, for the cyclic numbers of

EFF the ratios are 0.33, 0.51 and 1.32. The von Neumann ratio is significantly different from 2.00 at any reasonable level for Eras 1 and 2 and falls in the inconclusive region at level 0.10 for Era 3 [3]. Not surprisingly a great deal of autocorrelation was introduced by the moving average procedure.

4. Building the Models. For each data set the periodogram was produced using the fast Fourier transform. This required first subtracting the mean from each data point and extending the data with the appropriate number of zeros. The Fourier frequency with largest percent of the sum of ordinates was adjusted using the Brent iterative procedure. The data was then changed to be the residuals obtained by deleting this new frequency. The periodogram of these residuals was produced and the Fourier frequency with the largest percent of the sum of ordinates was chosen. The previously found (adjusted) frequency and this second frequency were fit at the same time by the Brent procedure to the original data. The two new frequencies were deleted from the data. This procedure was continued until one of our stopping conditions was reached.

Our stopping conditions were (1) five frequencies having been fit, (2) a von Neumann ratio for the residuals being sufficiently different from two for the eras [4], and (3) the periodogram of the residuals being just that of noise.

The data was linearly detrended if a period longer than the data set emerged in the original periodogram.

A number of components beyond those presented herein were investigated. By overfitting the data more confidence was gained regarding the choice of model[4]. One symptom of overfitting is the failure of the Brent procedure to converge with even as many as one hundred iterations. Usually only a few iterations were necessary.

Most of our computer programs are adaptations of those found in Bloomfield [2]. We chose not to taper the data since leakage did not affect our procedure.

This is a fitting exercise rather than smoothing. This type of fit has the advantage of linearity in the sum of the sinusoidal terms. One frequency can be considered at a time so that its impact can be weighed separately. The procedure and the programs used show high resolution, that is, frequencies close together can be distinguished and remain stable as more frequencies are introduced.

5. Model for All 94 Congresses. The mean of all 94 Congress' NPL is 430.6 laws and the sum of squares is 7.93×10^6 (standard error 294). The non-circular von Neumann ratio of lag one is 0.22 which indicates autocorrelation. However, this can be traced to the use of an overall mean which forces the denominator to be large. Each era has little autocorrelation as shown in Sections 3 and 7. The linear least-squares fit has a von Neumann ratio of 0.94 and its correlation coefficient is 0.877.

The model developed by our procedure is

$$\begin{aligned} NPL = {} & 9.39\,N - 20.25 + 60.38\cos(.084(N-1)) \\ & - 35.46\sin(.084(N-1)) \\ & -45.09\cos(.197(N-1)) + 55.82\sin(.197(N-1)) \\ & -41.37\cos(.396(N-1)) + 35.30\sin(.396(N-1)) \\ & +10.29\cos(.673(N-1)) + 53.18\sin(.673(N-1)) \\ & +27.21\cos(1.510(N-1)) - 33.95\sin(1.510(N-1)). \end{aligned}$$

N is the Congress number. Its residual sum of squares is 0.93×10^6 (standard error 100). The von Neumann ratio is 1.61.

It is interesting to examine each frequency separately. An $\omega = 0.084$ radians corresponds to a period of 74.8 Congresses. This is simply a modification of the linear trend. An $\omega = 0.197$ corresponds to a 31.9 Congress period and is a rough indication of the eras. These two frequencies together comprise forty eight percent of the sum of amplitudes of the five components.

The remaining three components have periods of 15.9, 9.3 and 4.2 Congresses. In a five Congress moving average smoothing their amplitudes would be multiplied by 0.85, 0.60 and 0.17 respectively.

Evaluating this model at N=95 and 96 for a forecast of NPL for the present and the next Congress yields the estimates 750 and 840 laws with standard error of 100. That is, the current down swing in NPL will be reversed.

6. Models for Each EFF Era. Models for each EFF era and some of their statistics are displayed in Table 1. The periods range from 2.5 to 24.2 Congresses. The models are graphed in Figure 2.

Since the analysis in EFF of the relationship between certain political variables as independent variables and NPL as dependent variable was performed using linear regression techniques (see Section 1), column 1 of Table 2 suggests that the strengths of the relationships may be different if NPL rather than cyclic numbers were used. However, the r's for Eras 1 and 2 are significantly different from zero at the 0.02 level. The unstarred r's in Table 2 are not significant at the 0.05 level. Of course, using the wider eras described above would also affect these.

The second column of Table 2 displays r's which measure the degree of fitting of the models. The residual sum of squares of Table 1 does also.

The third column of Table 2 is interesting. Since the residuals of the models in this report are produced by fitting and the cyclic numbers (residuals) of EFF are yielded partly by a smoothing operation, the correlations would not necessarily be high. However, in Era 3 there is a period of 4.9 Congresses (ω=1.286) which is destroyed by the EFF five Congress moving average. Hence, this was also a fit in a sence.

No forecast is possible for the 94th Congress since Era 3 stops at the 85th Congress.

Figure 1 is a graph of this model. The choices of cut off points at Congresses 36-37 and 66-67 of the extended eras in Section 2 are reinforced by this model. The same eras are perceived by Fourier analysis as were chosen by the simpler criterium of the number of standard deviations from local means (Section 2).

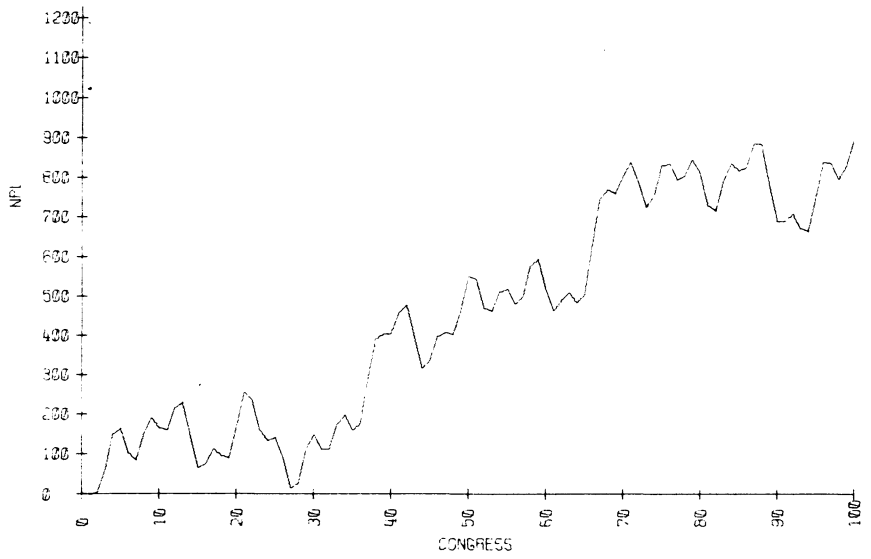


FIGURE 1: MODEL OF NPL FOR ALL 94 CONGRESSES (PLOTTED THROUGH THE 100TH CONGRESS)

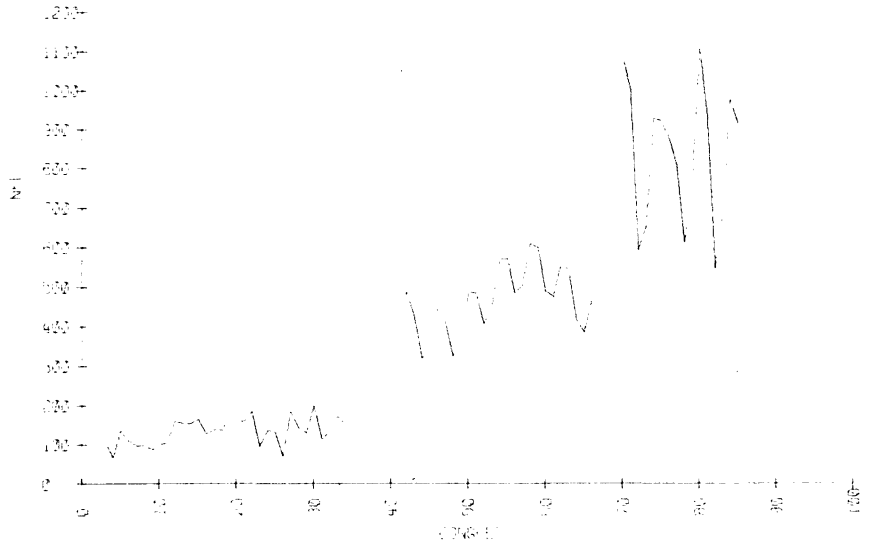


FIGURE 2: MODELS OF NPL FOR EACH EFF ERA

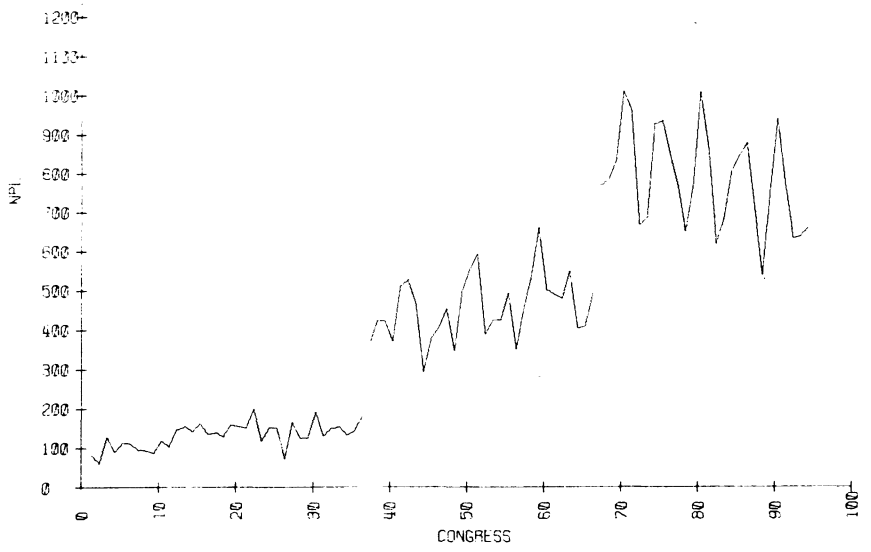


FIGURE 3: MODELS OF NPL FOR EACH FULL ERA

Era	Congresses	NPL				Model			Cyclic No.		
		Mean	Sum of Squares	Std. Error	von Neumann	Residual Sum of Squares	Std. Error	von Neumann	Residual Sum of Squares	Std. Error	von Neumann
1	3-34	134.3	4.31×10^4	38	1.55	0.63×10^4	14	1.56	0.76×10^4	16	0.33
2	41-66	464.7	2.98×10^5	111	1.42	1.35×10^5	75	2.18	0.55×10^5	48	0.51
3	70-85	839.8	5.77×10^5	203	1.75	1.02×10^5	85	1.60	0.20×10^5	37	1.32

Models Era 1: NPL = 1.84 N + 104.17 - 14.54 cox (.294(N-3)) - 15.04 sin (.294(N-3)) - 9.14 cos (.816(N-3))
+ 19.92 sin (.816(N-3)) + 17.63 cos (2.081(N-3)) - 0.51 sin (2.081(N-3))
+ 6.29 cos (2.270(N-3)) - 17.90 sin (2.270(N-3)) - 4.42 cos (2.527(N-3))
- 18.07 sin (2.527(N-3))

Era 2: NPL = 466.62 - 26.55 cos (.260(N-41)) - 84.48 sin (.260(N-41)) - 22.17 cos (1.547(N-41))
+ 69.77 sin (1.547(N-41))

Era 3: NPL = 821.43 + 179.48 cos (1.286(N-70)) + 81.80 sin (1.286(N-70)) + 84.53 cos (1.898(N-70))
+ 87.57 sin (1.898(N-70))

N is the Congress number.

TABLE 1: MODELS AND STATISTICS FOR THE EFF ERAS

Era	Congresses	NPL				Model		
		Mean	Sum of Squares	Std. Error	von Neumann Ratio	Residual Sum of Squares	Std. Error	von Neumann Ratio
1	1-36	132.1	4.90×10^4	38	1.45	1.22×10^4	19	1.64
2	37-66	456.8	3.14×10^5	106	1.41	0.99×10^5	60	2.09
3	67-94	786.3	8.00×10^5	175	1.39	3.70×10^5	119	1.57

Models Era 1: NPL = 1.79 N + 103.50 - 8.49 cos (.262(N-1)) - 16.34 sin (.262(N-1)) - 16.48 cos (.745(N-1))
+ 5.31 sin (.745(N-1)) + 12.72 cos (2.128(N-1)) - 10.14 sin (2.128(N-1)) - 3.26 cos (2.314(N-1))
- 15.52 sin (2.314(N-1)) - 8.62 cos (2.572(N-1)) - 16.20 sin (2.572(N-1))

Era 2: NPL = 4.33 N + 236.81 - 67.20 cos (.686(N-37)) + 26.99 sin (.686(N-37)) + 23.39 cos (1.506(N-37))
+ 70.99 sin (1.506(N-37)) + 11.45 cos (3.217(N-37)) + 33.07 sin (3.217(N-37))

Era 3: NPL = -5.58 N + 1243.12 - 123.90 cos (1.242(N-67)) - 90.88 sin (1.242(N-67)) + 26.31 cos (1.990(N-67))
+ 63.46 sin (1.990(N-67))

N is the Congress number.

TABLE 3: MODELS AND STATISTICS FOR THE EXTENDED ERAS

	Paired Models		
	NPL & EFF Cyclic No.	NPL & Residuals of the Models	EFF Cyclic No. & Residuals of Models
Era 1 (n=32):	0.424*	0.442*	0.181
Era 2 (n=26):	0.501*	0.672*	0.369
Era 3 (n=16):	0.135	0.420	0.661*

TABLE 2: CORRELATION COEFFICIENTS AMONG THREE REPRESENTATIONS OF NPL (*means $\alpha \leq 0.02$)

7. Models for Each Extended Era. Models for each extended era and some of their statistics are displayed in Table 3. Figure 3 is a graph of all three eras. A comparison of Figures 2 and 3 shows that nearly every peak and valley in the two graphs correspond. Hence, the shorter EFF eras are indeed subsets of the full eras.

In Era 1 there are two longer periods of 24.0 and 8.4 Congresses. Era 2 possesses two large periods of 9.2 and 4.2 Congresses. The longer period in Era 3 is 5.0 Congresses. The remaining era periods are fewer than 3.2 Congresses. Each era can be characterized not only by its level of productivity but also by its set of frequencies or periods. Each of these longer periods arise approximately in the model of all 94 Congresses in Section 5.

Evaluating the model for Era 3 at N = 95 & 96 for a forecast of NPL for the next two Congresses yields the estimates 824 and 878 laws with standard error 119. Hence, this model also predicts an upswing in NPL.

8. Summary. A model was constructed by Fourier decomposition of the time series of NPL for all 94 Congresses. Eras of the level of productivity of public laws were discerned. The three eras are Congresses 1-36, Congresses 37-66, and Congresses 67-94. The three eras of a previous study (EFF) are subsets of these eras.

Models were constructed for each EFF era and each of the full eras.

The model for all 94 Congresses and the model for the most recent full era predict a reversal in the present downtrend in NPL.

References

[1] Eilenstine,D.L., Farnsworth,D.L. and Fleming,J.S. (1977). "Trends and Cycles in the Legislative Productivity of the United States Congress, 1789 - 1976," Quality and Quantity 11, to appear in December 1977.

[2] Bloomfield,P. (1976). Fourier Analysis of Time Series: An Introduction. J. Wiley and Sons, New York.

[3] Yamane, T. (1973). Statistics, An Introductory Analysis. Third Edition. Harper & Row, New York.

[4] Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control. Revised Edition. Holden-Day, San Francisco.