I

DEVELOPMENTS IN SOCIAL STATISTICS

Chairman, Harold F. Dorn, National Institutes of Health

Page

Race, Economic Attitudes, and Behavior - James Morgan, University of Michigan and Martin David, University of Wisconsin..... 2

The Negro as an Immigrant Group: Recent Trends in Race and Ethnic Segregation in Chicago - Karl E. Taeuber and Alma F. Taeuber, University of Chicago..... 11

Studies of Comparative Prevalence of Cardio-Respiratory Disease in England and the United States - Walter W. Holland, London School of Hygiene and Tropical Medicine and Johns Hopkins University..... 19

RACE, ECONOMIC ATTITUDES, AND BEHAVIOR

By James Morgan, Survey Research Center,
University of Michigan and Martin David,
SSRI, University of Wisconsin

All of us have made casual observations on the differences between white and Negro families. Some have gone further and have scientifically described differences in some detail, partly because skin color is such an obvious and easily identified demographic characteristic, partly because the history of the Negro population and its economic advances have directly and indirectly shaped some of the major social issues of this generation, and partly because prejudice against non-whites has been one of the principal areas of myopia in the democratic process in this country. In this paper we add a few recent observations on white non-white differences to the many observations that have been made in the past. Our primary effort, however, is to examine those differences to determine to what extent observed differences are correlated with factors other than race. In the process we gain some insight into the mechanisms whereby social and economic differences between white and non-white groups are sustained. The study thus lends some perspective on the extent to which dynamic changes in our economy in the next decade or so will be translated into social change and economic improvement for non-white groups.

The problem we are concerned with, then, is disclosing the chain of causation that leads to racial differences. Causation is a complex process, and the direction of influence is sometimes difficult to untangle. Such factors as parental background and race, perhaps religion, come earlier in the life history of an individual than other factors such as the amount of education completed, the kind of occupation and place of residence selected, and the individual's current attitudes and behavior.

We can think of a sequence of causation, starting with race as a determinant of the life history of a man, a determinant which is inextricably fixed at birth. Race may have already affected other parental conditions such as the father's religion, place of residence, income, family size, and attitudes. These family background factors in turn, along with the individual's race, affect his own educational achievement, occupation, mobility, level of achievement motivation, and family planning. The individual's achievements and past decisions along with everything else already determined can be thought of as affecting the individual's present attitudes and behavior -- his belief that hard work pays off, his attitude toward education of his children, his attitudes toward responsibility for relatives and toward living with relatives and toward government responsibilities for the unemployed, students, and the aged.

Finally, all these aspects of attitudes and behavior affect the individual's present condition: hourly earnings, hours of work, accumulation of savings, insurance coverage, education of his children, and the extent of his planning for the future.

It is the present condition in which we are generally interested. Yet from the above argument it is obvious that while we could attribute it all to race, as the only thing that has been fixed and immutable from the beginning, it is more important to see the mechanisms through which race affects behavior, particularly if one wants to ask the question whether racial differences can be reduced or eliminated.

With this framework as background, let us turn now to some descriptive statistics,

first, to relate race to other parental background factors
second, to relate race to past decisions and actions which can no longer be changed
third, to relate race to present attitudes and behavior
and finally, to relate race to the present situation of individuals and their plans for the future.

Thus we examine a set of statistics that follow the life history of an individual from the time he is born, either white or non-white, through the time he is educated, enters the labor market, starts a family of his own, down to the present time, his life situation, and his planning for the future.

In any cross section, we have representation from several generations, and in a changing world, it may be important to distinguish one generation from the next. For this reason, we sometimes analyze separately those under 45 years of age, those 45 - 64 and those 65 and older.

In examining these statistics we have sometimes estimated the net effect of racial differences on behavior by explaining the dependent variable in a multi-variate regression that includes race and a relatively comprehensive set of other relevant independent variables. The net effect represents a standardized difference between the two groups, assuming other relevant influences on behavior are the same for the two groups. It is a factor which tells us how to make the best estimates of economic behavior from data available in the single cross-section of the U. S. on which we based this study. Estimates of the net effect are not presented as absolute measures of racial differences.

There are important influences on behavior which were not included in the analyses that represent a more basic cause of differences between the groups. Nevertheless the difference between the actual observed values of dependent variables for white and non-white groups and the values that are

predicted by using the net effect, gives an idea of the extent to which racial differences can be assigned to other factors. For example, it is quite revealing that educational aspirations of non-whites for their boys are fundamentally the same as whites at the same socio-economic level. The actual expectations of non-whites are lower than whites primarily, we would infer from our multi-variate analyses, because of factors other than skin color.

Insofar as white, non-white differences are larger than the standardized net effects, the analyses suggest that one may reduce actual differences between the two groups by creating an environment in which non-whites can achieve the same education, occupation, and geographic location as whites. The net effect represents a difference between the two groups that is as yet unanalyzed -- some part may be the consequence of discrimination, some part the result of personality differences, and some part may be associated tastes and cultural differences. We make no attempt to evaluate the origins of the net effects which must be further analyzed to determine basic causes.

Association of Race with other Background Factors

We need not present much data to document what is well known: that most non-whites come from large families in the rural South where the parents had little education and unskilled occupations. Table 1 shows that more non-whites come from families where they had many brothers and sisters.[1] More people than one might think come from large families because the larger the family, the more siblings there are who can fall into a sample of the children. In any case the limited economic resources of the unskilled Negro had to provide for more children so that investment in the education of children was more difficult for the non-white family.

The trend in the proportion of younger heads who come from large families indicates that the family size of whites has declined more rapidly than non-whites.[2]

[1]All the data in this paper come from a national probability sample, oversampled among lower income units of earning age, and weighted. The main analysis of the study appears in Morgan, David, Cohen, and Brazer, Income and Welfare in the United States (New York: McGraw-Hill Book Company, 1962). There is a slight upward bias in parental family size for non-whites because our base is spending unit heads, not all adults, and more non-white females are spending unit heads.

[2]As the length of a generation is shorter among non-whites than among whites, the finding understates the differential rate of propagation of the two groups.

A substantial fraction of both groups come from farm backgrounds, but a much larger proportion of non-whites report that their father was an unskilled laborer or service worker (39 per cent non-whites versus 11 per cent white). And, of course, more of the non-whites grew up in the South. The proportions of each group who grew up on a farm in the Deep South and had two or more siblings are 13 per cent for whites and 41 per cent for non-whites.

Table 1 also provides a measure of achievement motivation, which is thought to be a stable personality disposition resulting from early childhood training experiences. It can be thought of as a crude substitute for direct measures of parental attitudes and training. Non-whites scored consistently less than whites on the achievement index. However, the particular measure used may lead to spuriously low non-white scores, because it is based on the extent to which individuals differentiate between occupations requiring greater and lesser degrees of skill and education; non-whites have less experience with some of the occupations rated than whites.

Past Decisions and Experience of Spending Unit Heads

Non-whites generally complete less formal education than whites. Twenty-three per cent of non-whites not only did not go to college, but were three or more years behind their usual grade in school when they left school, as compared with 10 per cent of whites. While this finding appears unusual, it confirms a pattern of erratic school attendance of Negro children demonstrated in Census data.[3]

When one takes account of the lower education of non-white fathers and other depressing background factors, the difference in educational achievement is reduced to less than one fourth of its unadjusted size. Even the net difference, however, is significant and hence not completely explained by the parental factors we were able to measure, including the achievement motivation inculcated by early training.

The first three lines of Table 2 indicate that educational differences are narrowing. The measures of performance and achievements are crude, and take no account of the quality of education. As there is reason to believe that quality differentials are also narrowing, younger non-whites probably received far more adequate education than their parents to an extent greater

[3]E. Bernert, America's Children (New York: John Wiley & Sons, Inc., p. 76).

than the table suggests.

The next four lines of Table 2 indicate that non-whites have less occupational mobility but more geographic mobility off the farm and out of the Deep South. The restricted occupational mobility has many causes, including the lower level of education of non-whites. The differences are more dramatic when one remembers that many whites start in high-status occupations and do not need to move up. Mobility in a more general sense as measured by whether the person has lived in three or more states since his first job is, if anything, lower for non-whites.

Another crucial aspect of past behavior is the family planning decisions that are made. Non-whites tend to start their families earlier and to have more children, both of which reduce their ability to stay in control of their own finances and to provide for their children. The last two lines of Table 2 indicate that these differences may be increasing in importance.[4]

One result of this and the other background influences, is that present day non-white spending units have provided less education for their children than have the whites, by more than 2 complete grades. However, in a multivariate analysis which simultaneously takes account of a number of other factors (the father's education, the difference in education between father and mother, difference in education between father and grandfather, number of living children, father's occupation, achievement motivation, religion, attitude toward hard work, age at time of birth of eldest child, age, and several mobility factors), the net difference is reduced to slightly more than one half a grade. The difference is significant, but small in comparison with variation in educational achievement associated with the other factors studied.

Present Attitudes

Parental background, and the individual's own past history affect his present situation partly through their effects on his attitudes. Table 3 presents some differences in attitudes. Non-whites are somewhat more likely to believe that luck or help from friends determine a man's success, though the majority still feel that it is hard work that matters most. The differential predisposition to attribute success to luck may partly explain, or be explained by, the significant difference in the number of hours worked by white and non-white heads, a difference that persists even after adjustments have been made for occupation, education, and so forth.

Planning ahead is crucial to economic success and stability. Non-whites reported that they felt less able to plan ahead, and out of six possible places in the questionnaire where they could have indicated actual plans, were less likely to have any explicit plans. (Table 3, lines 2, 3, and 4.) While low income groups generally had less plans than others because of unstable employment and other factors somewhat beyond their control, lack of the particular characteristics taken as evidences of planning indicates that the family has few or no resources to meet emergencies such as sickness or death of the breadwinner. Hence planning becomes an important mechanism through which race creates economic differentials.

Non-whites are much more in favor of having the government, rather than relatives support the aged. (Table 3, last line). Actually a multivariate analysis which takes account of income and other factors makes the racial differences even greater.

On the other hand, non-whites have the most favorable attitudes of any subgroups examined toward older people living with their children or relatives. One fourth favor it without qualification, whereas the overwhelming majority of other groups are opposed. However, non-whites are no more likely than whites actually to live with relatives once we allow for low income and other pressures on them to do so. These attitudes may determine the actual living arrangements of non-whites. Non-whites are more likely to live with relatives, in fact, yet may have less of a desire to do so, ceteris paribus. The attitude that government should support the aged is consistent with unwillingness to live or move in with relatives, while children's acceptance of extending their family to include dependent parents may explain why as many non-whites as whites support relatives according to our standardized estimates. (See the section on living arrangements that follows below.)

Non-whites are also more in favor of government action on two other fronts. They were much more likely to say that there should be more tax support for colleges, 75 per cent as against 49 per cent of whites. And they were more likely to favor such support for all students, or those with need, rather than those with ability. Twenty-nine per cent of whites mentioned ability, alone or in combination with need, but only 15 per cent of the non-whites mentioned it.

The third area of government studied was the

[4] However the increase in early births among non-whites is not supported by Current Population Report Statistics which indicate no trend in the first births per 1000 women occurring before age 20 (Bureau of the Census, Current Population Reports, P-20, #108, July 1961, Table 8). The trend shown in Table 2 persists when data are tabulated for spending units headed by a male. Census data from the same report indicate significantly earlier births among non-white women in all cohorts than among white women.

level of unemployment compensation payments. When asked "Do you think unemployment compensation payments should be higher, lower, or the same as they are now?", non-whites were significantly more likely to say "higher" even after adjusting in a multivariate analysis for a number of other factors such as religion, unemployment in the state and in the family, labor union membership, family composition, income, unemployment experience, and so forth.

Present Situation, Behavior and Future Plans, and the Extent to Which They Result from Measured Intervening Variables

We have seen that non-whites are less likely to plan ahead. Perhaps the most crucial planning, from a long-range point of view, is the planning of the education of children. Here, race is associated with differential expectations in the case of girls, but not of boys. The reason is that the depressing effects of low father's education, low income, etc., operate about as expected in the case of plans for non-white boys, but do not appear to depress expectations for the non-white girls. While the differences in expectations among whites hinge largely on whether a college education is expected, among non-whites they hinge on whether finishing high school is expected. The high expectations among non-whites for their girls reflect an insistence that the girls finish high school. Given the increased occupational opportunity for secretarial work rather than housework that a high school diploma makes possible for non-white girls, the attitude seems reasonable.

Housing and Living Arrangements

We have seen that non-whites have larger families. Table 4 indicates that non-whites are less likely to be married, and those that have children are less likely to be living with their spouses.

On the average, non-whites are 6 per cent more likely to live with relatives, or 7 per cent more likely to provide housing for relatives. Adjustment for other factors such as income, age, stage of the family life cycle, and number of children makes the first relationship significantly reversed, and the second non-significant. In other words, after accounting for other influences, non-whites are actually less likely to live with relatives. More non-whites are without a spouse. (Table 4) However, this condition is less likely to cause them to double up than it is among the whites. This finding contradicts the common notion that non-white families have a greater propensity than white families to support dependents, such as uncles, parents, and cousins, by doubling up.

Non-whites are less likely to own their own home than whites. A multivariate analysis which incorporated age, spending unit income, number of persons in the unit, whether income last year was unusual, number of major earners, and education of the head of the unit, showed that much of the 18 per cent difference in ownership by race was attributable to the other factors, but a significant 8 per cent difference remained. Presumably discrimination, income uncertainty, and FHA regulations which frequently do not allow credit for the wife's income all help create this difference.

A similar analysis of house values for home owners revealed that adjusting for other factors than race reduced the difference in house value from $5,700 to an insignificant minus $1,500. This finding does not offer any clues to quality of housing or its adequacy, but merely indicates that the market value of homes owned by whites is $1,500 less than the market value of houses owned by Negroes in similar circumstances.

Property taxes for home owners averaged $106 for non-whites and $183 for whites. On the other hand, if one imputes a property tax payment to renters, non-white renters appear to pay more taxes (and more rent) at each level of income or welfare, than white renters. The combined effects of the lower incomes, and the other differences, leaves all non-whites paying an average property tax of $67 as compared with an average of $130 for whites.

The difference between owners and renters may be the result of discrimination. If non-whites by being excluded from many areas, pay higher prices for equivalent houses, and assessors do not revise assessments upwards in non-white areas because of this, then non-whites will be buying smaller houses than whites at the same income level and will be paying lower property taxes relative to their incomes (not relative to house values). At the same time, non-white renters will be forced to pay higher rents (implying a higher estimated property tax by virtue of our imputation procedure) for our equivalent housing.

Finally, since non-whites have more children, non-whites get more public school benefits than whites, $310 per family compared with $200 for whites. However, non-white families tend to live in counties where average expenditures per pupil on primary and secondary education are slightly less than expenditures in counties more predominately white ($320 for whites, $301 for non-whites).

Other Aspects of the Present Situation

Table 5 shows that the concentration of non-whites in the South is greater among the older generation, that unemployment is more serious among non-whites, and that non-whites, particularly the younger ones, are more likely to live in large cities.

Employment and Earnings

Differences in economic position as between whites and non-whites are a complex result of differences in the earning rate and hours of work of the heads of units, and of wives, combined with the fact that significantly fewer

non-white units contain a wife. Over-all, non-white units have annual earnings 40 per cent lower than white units. Non-whites receive less per hour and work fewer hours. Their wives are much more likely to work, but earn less per hour and work fewer hours than white working wives. Indeed, the increased proportion of non-white wives working is almost offset by the fact that fewer units have wives. Thirty-five per cent of all non-white spending units have a working wife compared with thirty-one per cent of all white spending units.

To what extent can these differences be attributed to differences in education, occupation, age, place of residence, and other intermediate variables more subject to public policy influences than race itself? The adjusted differences in Table 6 indicate that if whites and non-whites were alike in other respects (things we could measure like education, age, occupation) the differences in both wages and hours of men, and wages of women would be a great deal smaller.

The difference in earnings of white and non-white spending unit heads averages $1,750. The differential is attenuated to $840 by the multivariate adjustment. Similarly, in spending units where both the head and wife are present the difference in actual average earnings of head and wife is attenuated from $2,048 to $981. While the reduction in both differentials is substantial, a highly significant differential remains. It remains for some future study to determine what part of this net difference in earnings can be explained by overt discrimination against non-whites and what part can be explained by other dimensions not studied here.[5]

Interestingly enough, the analysis indicates that labor force participation of the wife and her hours would continue to differ between whites and non-whites even if they were similar as to formal education, age, family status, and so forth.[6] This difference is corroborated by a difference in labor force experience of the wife. The average non-white wife has worked more than ten years altogether (10.2) and the average white only seven (7.3), including work before marriage. If one takes account of other factors, such as age, education, attitude of the head about wives working, etc., the difference is increased to nearly three and one-half years, a highly significant difference.[7]

[5]Some of the differential in earnings which remains may be explained by factors not included in the multivariate analysis or more precise measures of the characteristics studied.

[6]Among those with school-age children, 61 per cent of non-white wives worked as compared with 40 per cent of white wives.

[7]The question was: "How many years has she worked (did she work) altogether (including years that she worked before she got married)?"

Another insight into the differential earnings of whites and non-whites comes from an analysis of those who reported some disability. We found no differences between the extent to which white and non-whites reported that the disability limited their ability to work. One might have anticipated that disabled non-whites would find it more difficult to get a job they could handle or would use physical incapacity as a crutch to excuse limited work effort; yet the data give no support to either hypothesis.

Other Work

The effects of race on consumption and saving behavior has been carefully studied, particularly by Klein and Mooney.[8]

The effects of discrimination through occupational restriction have been documented and discussed by Gary Becker.[9] The greater impact of unemployment and the failure to share in income gains since 1950 are shown in U. S. Department of Labor, The Economic Situation of Negroes in the United States, Bulletin S3, October, 1960.

Summary

Clearly, the relatively deprived background of non-whites has created obstacles which have made it difficult for the non-white family to advance economically and socially in our advanced industrial society. Lack of education on the part of parents has perpetuated itself in lack of education of children, and this in turn has limited employment and has made it more difficult for non-whites to protect themselves against contingencies of illness and other economic catastrophes. Lack of the same degree of family planning prevalent among white families has also contributed to their economic difficulties.

At the same time the non-white family has developed some interesting traits which may serve to reduce the situational poverty that surrounds them. Their aspirations for educating girls are relatively higher than the aspirations of whites. The non-white wife tends to remain in the labor force for a longer period than the wife in a white spending unit, although she works fewer hours.

Insofar as the economic disadvantages of the non-whites operate through lower levels of

[8]L. R. Klein and H. W. Mooney, "Negro-White Savings Differentials and the Consumption Function Problem," Econometrica, 21 (July, 1953), 435-456; see also Marcus Alexis, "Some Negro-White Differences in Consumption," American Journal of Economics and Sociology, 21 (January, 1962), 11-28.

[9]Gary S. Becker, The Economics of Discrimination, Chicago, 1957; see also comment by Alton Rayack in May, 1961 Review of Economics and Statistics, and rejoinder by Becker in May, 1962 of that journal.

education and restricted job opportunities, public policy directed toward more equal opportunity can focus directly on opportunities for education and employment. Insofar as the racial differences are perpetuated by differences in attitude and belief about the world, it will be difficult to change them until we change the situations which justify those attitudes and beliefs. But the clearer it becomes that hard work may pay off, that planning ahead is important and possible, the more likely attitudes are to change too.

Table 1

Background Measures for Three Generations of Whites and Non-Whites
(per cent of each age-race group)

	Age					
	18 - 44		45 - 64		65 or older	
Per cent of heads of each group who:	White	Non-White	White	Non-White	White	Non-White
Have four or more siblings	45 <	63	58	63	69	88
Have high index of need for achievement[1]	33	26	31 >	18	27	19
Number of cases	1300	222	931	147	349	48
Per cent of sample	46	5	31	3	13	2

Source: Survey Research Center, Study 678

Differences marked by inequality signs are statistically significant at the 5 per cent level by a conservative estimate allowing for sample clustering.

[1] It should be kept in mind that the data refer only to heads of adult units (adult individuals or couples), hence are mostly men. There are difficulties in assuring comparability of measures as between men and women. However, it is interesting that a national probability sample of adults using a TAT measure of need-achievement found much greater differences between Negro and white women than between Negro and white men. It was the Negro women who had the smallest proportion with high need-achievement scores.

See Joseph Veroff, John W. Atkinson, Sheila C. Feld, and Gerald Gurin, "The Use of Thematic Apperception to Assess Motivation in a Nationwide Interview Study," Psychological Monographs, 74, No. 499, 1960.

Table 2

Past Decisions and Experience of Three Generations of Whites and Non-Whites
(per cent of each age-race group)[1]

Per cent of heads of each group who:		Age 18 - 44 White		Age 18 - 44 Non-White		Age 45 - 64 White		Age 45 - 64 Non-White		Age 65 and older White		Age 65 and older Non-White
Education	Report grades in school were above average	30	>	20		27		20		25		15
	Report grades not above average and were a year or more behind age group when dropped out of school	11	<	24		21	<	38		26		26
	Report grades not above average, and three or more years behind when dropped out	2	<	8		5	<	20		13		17
Mobility	Started in a middle or low status job and moved up to a higher status job	30		22		34	>	20		22		20
	Grew up in a rural area but now live in a city[2]	41	<	65		45		60		43	<	75
	Grew up in the Deep South but now live entirely outside the South[3]	15	<	34		13	<	33		7	<	39
	Have lived in three or more states since first job	10		8		19		16		20		15
Family planning	Report that head was under 20 when first child was born	6	<	20		4		8		6		8
	Have five or more children	6	<	19		10	<	24		17		28

[1] See Table 1 for number of cases; differences marked by inequality signs are statistically significant.

[2] Per cent of heads who grew up in rural areas

[3] Per cent of heads who grew up in the Deep South

Table 3

Present Attitudes of Three Generations of Whites and Non-Whites
(for all heads of spending units)

Per cent of heads of each group who:	Age 18 - 44 White		Age 18 - 44 Non-White		Age 45 - 64 White		Age 45 - 64 Non-white		Age 65 and older White		Age 65 and older Non-White
Believe that hard work is more important than luck or help from friends in getting ahead	84	>	65		80	>	67		83		74
Feel able to plan ahead[1]	63	>	50		54	>	33		39		28
Give some evidence of planning, in any of six possible places1	93	>	85		94	>	75		81	>	41
Give 2 or more indications of planning out of a possible six[1]	80	>	61		82	>	50		55	>	28
Believe that government rather than relatives should have primary responsibility for the aged	23	<	41		27	<	43		34		47
Number of cases	1300		222		931		147		349		48

Differences indicated by inequality signs are statistically significant

[1] There may be some tendency for the planning index to be biased downward for low income people including non-whites since it counts such things as having savings of $500 or more and have a pension other than social security.

Table 4

Stage in Life Cycle within Race
(for all adult units)

Stage in life cycle	All	White	Non-White
Under 45, no spouse, no children	14	14	17
Wife under 45, married, no children	7	7	6
Wife under 45, married, children under 6	21	22	20
Wife under 45, married, children 6 or older	11	11	6
Wife 45 or older, married, children under 6	1	0	2
Wife 45 or older, married, children 6 or older	6	6	5
Wife 45 or older, married, no children	17	18	9
Head 45 or older, no spouse, no children	18	18	22
No spouse, children	5	4	13
Total	100	100	100
Per cent with no spouse		36	52
Number of cases		2887	509

Table 5

Present Situation for Three Generations of Whites and Non-Whites
(for all spending unit heads)

	Age					
	18 - 44		45 - 64		65 or older	
Per cent of each group who:	White	Non-White	White	Non-White	White	Non-White
Live in the South	21 <	42	25 <	52	14 <	64
Are unemployed usually or occassionally	7	12	6	11	1	5
Live in a city of 50,000 or more	30 <	65	31 <	54	33	41

Differences indicated by inequality signs are statistically significant.

Table 6

White Non-White Differences in Labor Force Participation and Earnings, Absolute and Standardized as if the Two Populations were Identical on Many Other Characteristics such as Education, Occupation, Age, etc.

	Unadjusted estimates		Difference (White minus non-white)	Adjusted (standardized) Difference[1] (White minus non-white)
	Whites	Non-whites		
1. Whether head of spending unit worked in 1959	.86	.84	.02	-.01
2. Hourly earnings of head if worked	$2.37	$1.60	$ .77	$.31
3. Hours worked by head if worked	2114	1894	220	116
4. Annual earnings of head if worked	$5000	$3040	$1960	$1010
5. Annual earnings, average including nonworkers	$4300	$2550	$1750	$ 840
6. Whether a wife[2]	.825	.696	.129	?
7. If wife, whether worked in 1959	.37	.50	-.13	-.10
8. Hourly earnings of wife if worked	$1.77	$1.16	$.61	$.11
9. Hours wife worked if worked	1375	1097	278	278
10. Annual earnings of wife if worked	$2434	$1272	$1162	$ 520
11. Annual earnings of wife, if a wife	$ 898	$ 636	$ 262	$ 62
12. Annual earnings of wife, averaged over all spending units	$ 741	$ 443	$ 298	$ 141
13. Annual earnings per spending unit of head and wife	$5041	$2993	$2048	$ 981

Row 4 equals the product of rows 2 and 3
Row 5 equals the product of rows 1, 2, and 3
Row 10 equals the product of rows 8 and 9
Row 11 equals the product of rows 7, 8, and 9

[1] Adjusted differences in rows 1, 2, 3, 7, 8, and 9 are based on multivariate analyses

[2] Having no multivariate analysis of "whether there is a wife," we used the unadjusted proportions from row 6 also in deriving the "adjusted" estimates.

THE NEGRO AS AN IMMIGRANT GROUP:

RECENT TRENDS IN RACE AND ETHNIC SEGREGATION IN CHICAGO*

Karl E. Taeuber and Alma F. Taeuber

Population Research and Training Center, University of Chicago

The processes by which various immigrant groups have been absorbed into American society are complex, and have been studied from a variety of viewpoints. A distinctively sociological approach to the topic views assimilation as a process of dispersion of the members of the group throughout the social structure, and not solely as a psychological and cultural process. In the conventional account, the immigrants have initially settled in ethnic colonies in deteriorated central sections of large cities, and have found low-paid unskilled jobs. Assimilation consists in part of a process of social and economic advancement on the part of the immigrant group and their descendants, along with a decreasing residential concentration in ethnic colonies.

The large-scale migration of Negroes to Northern cities began during the first World War. In a very real sense Negroes served as a native-born substitute to fill in the labor gap created by the cessation of large-scale immigration from abroad, at first due to the war, and then as a result of newly imposed restrictions on immigration. Like the immigrants from abroad, the Negro migrants from the South moved to urban industrial centers where they filled the lowest occupational niches and rapidly developed a highly segregated pattern of residence.

Viewing the obvious analogies between the Northern urban Negro population and the European immigrant populations which preceded it, some sociologists have concluded that the Negroes will undergo a similar process of "assimilation," and that it is only a matter of time until social and economic progress is translated into their residential dispersion.[1] Other sociologists believe that the Negroes in Northern cities are not following the immigrant pattern of socio-economic advancement and residential dispersion, but rather that the second-generation urban Negroes are occupying the same relative position in the society as did their parents.[2]

The question of whether or not a Northern urban Negro population can fruitfully be viewed as an immigrant population, comparable to the immigrant populations of earlier decades with respect to the nature and speed of assimilation, is the underlying theme in our consideration of recent trends in race and ethnic segregation in Chicago.

One type of data indicative of the intergenerational assimilation of the immigrant groups is presented in the first six columns of Table 1. For each of the larger ethnic groups, data for 1950 are presented showing the average standing on three measures of socio-economic status, standardized for age, of the first generation (the foreign born white, FBW) and the second generation (native white of foreign or mixed parentage, NFMP). The nationality groups are split into "old," "new" and "newer" groups in an extension of the traditional system. On the average, comparing within the first or within the second generation, the "old" immigrant groups are the best off on these measures, the "new" groups are intermediate, and the "newer" groups are the worst off. It cannot be determined from these data to what extent the old immigrants are better off by virtue of their longer average period of residence in the U.S., and to what extent they may have been better off at their time of immigration than the newer immigrants were at the time of their move.

Comparisons between the first generation (FBW) and second generation (NFMP) are a more direct means for determining the extent of assimilation. Although it is not always specified carefully in the literature, most discussions of assimilation view it as an inter-generational process, rather than simply a process of upward adjustment through time in the status of the original immigrants. Emphasis is usually placed on the higher status and lesser residential segregation of the children relative to their parents. Comparisons of corresponding status measures for the first and second generations in Table 1 reveal the expected pattern of intergenerational advance. Although data of the type shown in Table 1 cannot be interpreted unambiguously, and do not suffice for specific measures of the degree or pace of inter-generational change, they are probably adequate indicators of the general direction of change.

Measures of the changing residential patterns of the immigrant groups are given in columns 7 - 9 of Table 1. The measure is an index of residential segregation between the total foreign stock (FBW + NFMP) of each nationality and the total native whites of native parentage (NWNP). The indexes were computed from the distribution of each group among the 75 Community Areas of the city of Chicago, for 1930 (the last previous census year which included information on the total foreign stock) and 1960.[3] The degree of residential segregation from the native population is highest for the "newer" immigrants, and lowest for the "old" immigrants. Between 1930 and 1960, most of the ethnic groups became less segregated from the native population. Only for England, Ireland, and Sweden did the indexes fail to

decline, and these were already at relatively low levels.[4]

The residential segregation between Negroes and NWNP was much greater than that between any of the ethnic groups and the NWNP. Furthermore, the small decline from 84 in 1930 to 82 in 1960 was less than for most of the ethnic groups. In every case, the residential segregation of the ethnic group from the Negro population is much greater than its segregation from the white population (NWNP). The lowest indexes for immigrant groups in comparison with Negroes occur for the two "newer" groups, Mexicans and Puerto Ricans. The high magnitude even of these two indexes indicates that these recent inmigrants to the city are not joining or replacing Negroes in the existing Negro areas, but are moving into separate ethnic colonies of their own. Lieberson has demonstrated that although prior to the great Negro migrations of World War I there were instances of ethnic groups being more segregated from native whites than were Negroes, since 1920 the general pattern has been for Negro residential segregation to be the highest.[5]

Data similar to those presented for the various ethnic groups are presented for whites and nonwhites in Table 2. For each of 4 measures reflecting socio-economic status, there was improvement in the status of the nonwhite (predominantly Negro) population between 1940 and 1960. For two of these measures, there was a definite narrowing of the differentials between whites and nonwhites. The indexes of residential segregation between whites and Negroes, in the top panel of the table, show minor fluctuations around an extremely high level, and give no indication of the decline anticipated on the basis of the socio-economic advancement of the Negro population. That this is not an atypical finding can be indicated by reference to other data showing a long term historical trend toward increasing residential segregation between whites and nonwhites. Increasing racial residential segregation was evident in most large cities of the United States between 1940 and 1950, while during the 1950's, Southern cities continued to increase in segregation and Northern cities generally registered modest declines.[6] In broad perspective, then, it appears that the historical trend toward improving socio-economic status of immigrant groups has gone hand in hand with decreasing residential segregation. In contrast, Negro residential segregation from whites has increased steadily over past decades until it has reached universally high levels in cities throughout the United States, despite significant advances in the socio-economic status of Negroes.

The pattern of decreasing residential concentration of immigrant groups and increasing residential concentration of Negroes is not what would have been expected from the fact that many nationality groups worked hard at maintaining the ethnic colonies, whereas most of the major Negro organizations strive for residential dispersal. Furthermore, there were declines in the residential concentration of the immigrant groups almost from the initial formation of the ethnic colonies, and this dispersion was going on during the periods of rapid increase in immigrant populations.[7] These observations tend to discredit the argument that a major barrier to residential dispersion of the Negro population in Chicago is its continuing rapid increase. However, the size of the Negro population and the magnitude of its annual increase are larger than for any single ethnic group in the past, and comparisons with smaller groups are not completely convincing. That rapid increase of Negro population does not necessarily lead to increasing residential segregation was demonstrated directly in the inter-city comparative study previously cited. There was no definite relationship between increase in Negro population and change in the value of the segregation index. Indeed, during the 1950-60 decade, there appeared to be some relationship in the opposite direction.[8]

More significant in accounting for the divergent trends in residential segregation may be the different urban contexts in which the immigrant and Negro populations found themselves. Comparing the residential locations of Italian-born and Polish-born in Chicago in 1899 and in 1920, Wallace observed:

> "...it can be seen that the areas of greatest dispersion, low proportion, and presumably of 'second' settlement for many immigrants were those which were not settled at all in 1899.
>
> "The implication of this fact is that the so-called 'assimilation' process was not reflected by the geographic dispersion of the immigrant populations into 'cosmopolitan American areas.' The dispersal was more directly related to an increase in housing alternatives as the city grew at the periphery."[9]

By the time the Negro concentrations were forming near the central areas of Chicago, the city was built up, and the urbanized area extended well beyond the present boundaries. Residential alternatives at a price Negroes could afford and in a sufficiently close-in location to permit inexpensive commuting were no longer available.

It has been suggested that considerable time is required for Negroes to make the transition from a "primitive folk culture" to "urbanism as a way of life."[10] Several types of data indicate that large and increasing proportions of the Negro urban population are city-born and raised. For instance, there is a rapidly decreasing color differential in the percentage of the Chicago population born in the state of Illinois. In 1960, 44 per cent of the native-born nonwhite residents of Chicago were born in Illinois, as contrasted to 66 per cent of the white population.[11] National estimates for 1958 showed that of all males aged 45-64 living in metropolitan places of 500,000 or more population, 65 per cent of the nonwhites as compared to 77 per cent of the whites had lived in this size of city for 20 years or longer.[12] Estimates of the components of growth of the nonwhite population

of Chicago indicate that between 1950 and 1960 natural increase was as important as net in-migration, and that natural increase will in the future account for rapidly increasing proportions of the growth of nonwhite population.[13]

Unfortunately there is inadequate knowledge of the specific length of time under specified conditions for the required cultural transformation to occur. Wallace's quoted observations indicate a significant degree of dispersal over time among the first generation immigrants. More usually, such processes are conceived as primarily inter-generational. That many of the "first generation" Negro migrants to Northern cities have lived there for 20 years and more and that in the younger adult ages there are sizable numbers of "second generation" urban Negroes suggests to us that there has been ample time for adjustment to urban living, at least for large proportions of the Negro population. It is also clear that if Northern Negroes remain inadequately educated for urban living and fail to participate fully in the urban economy, the "primitive folk culture" of the South can less and less be assigned responsibility, and Northern cities will be suffering from the neglect of their own human resources.

The "visibility" of Negroes due to skin color and other features which make the large majority of 2nd, 3rd, and later generation descendants readily identifiable as Negroes is often cited as a basic factor in explaining the distinctive position of Negroes in our society. It is exceedingly difficult to assess the significance of visibility. For instance, there is no other group which is strictly comparable to Negroes regarding every factor except visibility. It is not completely irrelevant, however, to note that nonwhite skin color, by itself, is not an insurmountable handicap in our society. The socio-economic status of the Japanese population of Chicago in 1950 substantially exceeded that of the Negro population, and their residential segregation from whites, although high, was considerably lower than that between Negroes and whites.[14] Unfortunately there are no trend data available on the characteristics of the Japanese in Chicago. A more appropriate Japanese population for comparison, however, is the much larger one in the San Francisco area. A recent study there affirmed that "ethnic colonies of Japanese are gone or rapidly going," and documented their rapid socio-economic advance.[15]

In the traditional immigrant pattern, the more recent immigrants displaced the older groups at the bottom socio-economic levels. How do the Negroes compare with the other "newer" immigrant groups, the Mexicans and the Puerto Ricans? The limited data now available suggest that the Negroes may soon be left alone at the bottom of the social and economic scale. We have already noted (from data in Table 1) that the "newer" groups were in 1950 of very low status compared to the other immigrant groups, and that their residential segregation from the native whites of native parentage was the highest of all the immigrant groups. For 1960, the distribution within Chicago of persons born in Puerto Rico is available separately from those born in the U.S. of Puerto Rican parentage. Thus it is possible to compute indexes of residential segregation for first and second generation Puerto Ricans. For Chicago in 1960, these index values were 68.4 for the first generation and 64.9 for the second generation, indicating that residential dispersion has already begun for the Puerto Ricans. This difference actually understates the amount of dispersion, since the second generation consists in large proportion of children still living with their first generation parents.

Selected socio-economic measures for the Puerto Rican and the nonwhite populations of the city of Chicago in 1960 are shown in Table 3. On every measure, the Puerto Rican population is less well off -- it is less educated, of lower income, more crowded, less likely to be home-owners, less well-housed, and lives in older buildings. Yet the index of residential segregation (computed with respect to NWNP) for Puerto Ricans is 67 as compared to 82 for Negroes.

Thus far we have been making comparisons between Negroes and immigrant groups. With respect to the relationship between socio-economic status and residential segregation, it is appropriate to pursue a more direct approach. Since Negroes are disproportionately represented in low status groups, it might be argued that on this basis alone we would expect some segregation between whites and Negroes.[16] To the extent that this is the case, future economic advances on the part of the Negro population should be translated into lowered residential segregation. Before presenting our approach to this problem, let us emphasize that the task of partialling out a component of racial segregation due to economic factors involves some difficult methodological problems, and no method is entirely satisfactory.[17] In an effort to make a rough assessment of the relative impact of patterns of economic segregation on patterns of racial residential segregation, we will consider a simplified model.

Basically, the approach involves an indirect standardization of the family income distributions for the 75 Community Areas of Chicago. The "rates" for the standardization were the percentage nonwhite in each income interval for the city of Chicago. For example, nonwhites constituted 44 per cent of all families in 1960 with an income below $1,000, 44 per cent of families with incomes of $1,000-1,999, 40 per cent for the interval $2,000-2,999, etc. This set of "rates" was then applied to the income distribution for each Community Area to obtain an "expected" number of nonwhite resident families on the basis of income. The total population of a Community Area minus the expected number of nonwhites is equal to the expected number of whites. Given this pair of expected numbers for each Community Area, we can compute an index of residential segregation between

expected nonwhites and expected whites. If income alone determined the residential locations of whites and nonwhites, this index is the expected amount of racial residential segregation. In 1950, the white-nonwhite residential segregation index expected on the basis of income was 11, compared to the actual segregation index of 79. Thus in 1950 income differentials can account for 11/79, or 14 per cent, of the observed racial segregation. In 1960, the expected segregation index was 10 and the actual 83, so that income differentials can account for only 12 per cent of the observed racial segregation. The slight decline from 14 to 12 per cent perhaps suggests that economic segregation is becoming a lesser component of total racial segregation.

Another investigator, studying ethnic groups as well as nonwhites, carried out a similar procedure, but used the rent and value distribution of occupied dwellings rather than the family income distribution to partial out economic factors in residential segregation. He also found that a negligible part of total residential segregation is accounted for by economic differentials. Furthermore, economic segregation was less a factor in the segregation of Negroes from whites than it was in the segregation of immigrant groups from native whites.[18]

The data just cited are one indication that it is not Negroes' inability to pay for housing that accounts for their residential segregation. In fact, in Chicago in 1960 Negroes paid as much as whites for housing, regardless of their lower incomes. Median rents for both groups were $88, but Negroes obtained much poorer housing for their money.[19] To a very real extent, there exists a separate housing market for Negroes in Chicago, so that their economic status cannot be used except in exceptional circumstances to obtain unsegregated housing. Regardless of their assimilation to urban living and their advancing economic position, therefore, Negroes have been unable to achieve the residential dispersion undergone by the second and third generation immigrant groups.

The judicious conclusion from our review of a variety of pieces of data is that we simply do not know enough about immigrant assimilation patterns and patterns of changing socio-economic status in the Negro population to be able to compare the two. With respect to immigrants, we are unable to find in the literature any satisfactory statistical specification of the processes involved. The available census data are of little use in this regard. For the Negro population, we have no data at all permitting inter-generational comparisons between migrants from the South and Negroes raised in Northern cities.[20] Thus any trends toward socio-economic advancement and residential dispersion on the part of "second generation" Negroes may be confounded in the data for the total Negro population.

If we can be allowed a brief moment of freedom from judicious interpretation of our data, we find ourselves in general agreement with the view that it is misleading to regard Negroes as another immigrant group. Even adopting a very simple formulation of assimilation as involving socio-economic advancement and residential dispersion, we do not think the data for Negroes can be interpreted as fitting the pattern. The second generation persons from several countries, in fact, are of higher socio-economic status than the total native whites of native parentage. Relatively few Negroes in Chicago have white collar jobs or have incomes above the median level for whites, and yet there are large numbers of adult Negroes who were born in the city. Basic differences between the Negroes and the immigrant groups seem to us implicit in the failure of residential desegregation to occur for Negroes, while it has continued to take place for the immigrant groups.

In view of the fundamental impact of residential segregation on extra-legal segregation of schools, hospitals, parks, stores, and numerous other facilities, the failure of residential dispersion to occur strikes us as an especially serious social problem. Although socio-economic advance and residential dispersion occurred simultaneously for the various immigrant groups, a causal relationship cannot be assigned. Nevertheless, it is apparent that the continued residential segregation of the Negro population will act as an impediment to the continued "assimilation" of Negroes into full and equal participation in the economy and the society at large.[21]

FOOTNOTES

*Paper No. 15 in the series, "Comparative Urban Research," issuing from the Population Research and Training Center, University of Chicago, under a grant from the Ford Foundation.

[1]Philip M. Hauser, "On the Impact of Urbanism on Social Organization, Human Nature and the Political Order," Confluence, VII (Spring, 1958), 65. Elsewhere Professor Hauser has expressed a more cautious view, emphasizing the lack of definitive knowledge; see his Population Perspectives (New Brunswick: Rutgers University Press, 1960), p. 129.

[2]Again we can cite a colleague, D.J. Bogue, "Chicago's Growing Population Problem," Commerce, 59 (July, 1962), 31.

[3]The index of residential segregation is an index of dissimilarity between the residential distributions of each group. It ranges from zero for an absence of segregation to 100 for complete residential segregation. For further discussion, see Otis Dudley Duncan and Beverly Duncan, "A Methodological Analysis of Segregation Indexes," American Sociological Review, 20 (April, 1955), 210-217.

[4]For a more detailed discussion of these patterns, using data for 1930 and 1950, see Otis Dudley Duncan and Stanley Lieberson, "Ethnic Segregation and Assimilation," American Journal of Sociology, LXIV (January, 1959),

364-374.

[5]Stanley Lieberson, "Comparative Segregation and Assimilation of Ethnic Groups," unpublished Ph.D. thesis, Department of Sociology, University of Chicago, 1960, p. 179.

[6]Karl E. Taeuber, "Negro Residential Segregation, 1940-1960: Changing Trends in the Large Cities of the United States," paper read at the Annual Meetings of the American Sociological Association, 1962.

[7]David A. Wallace, "Residential Concentration of Negroes in Chicago," unpublished Ph.D. thesis, Harvard University, 1953.

[8]Taeuber, op.cit.

[9]Wallace, op.cit., p. 205.

[10]Philip M. Hauser, "The Challenge of Metropolitan Growth," Urban Land, XVII (Dec., 1958), 5.

[11]Data from the 1960 Census, General Social and Economic Characteristics, Illinois, Tables 72 and 77. Foreign born were included in the denominator for whites, but ignored for nonwhites on the assumption that they were primarily Orientals rather than Negroes.

[12]Karl E. Taeuber, "Duration-of-Residence Analysis of Internal Migration in the United States," Milbank Memorial Fund Quarterly, XXXIX (Jan., 1961), Table 3.

[13]D. J. Bogue and D. P. Dandekar, Population Trends and Prospects for the Chicago-Northwestern Indiana Consolidated Metropolitan Area: 1960 to 1990 (Chicago: Population Research and Training Center, University of Chicago, 1962).

[14]Although the maximum value of the residential segregation index is less than 100 for ethnic groups of small size, this is not sufficient to vitiate the Negro - Japanese comparison.

[15]Harry H. L. Kitano, "Housing of Japanese-Americans in the San Francisco Bay Area," p.184, in Nathan Glazer and Davis McEntire, eds., Studies in Housing and Minority Groups (Berkeley and Los Angeles: University of California Press, 1960.)

[16]For a discussion of class residential segregation in Chicago, see Otis Dudley Duncan and Beverly Duncan, "Residential Distribution and Occupational Stratification," American Journal of Sociology, LX (March, 1955), 493-503.

[17]A general discussion of this problem can be found in the section on explanation of areal variation in Otis Dudley Duncan, Ray P. Cuzzort, and Beverly Duncan, Statistical Geography (Glencoe: Free Press, 1961).

[18]Lieberson, op.cit., p. 127.

[19]For a comprehensive analysis of white-nonwhite differentials in housing in Chicago as of 1956, see Beverly Duncan and Philip M.Hauser, Housing a Metropolis - Chicago (Glencoe: Free Press, 1960), Ch. 6.

[20]The only approach possible with census data is a comparison between recent migrants and the rest of the population, and then only the residential distributions are available, with no socio-economic characteristics. Using 1960 Community Area data for Chicago, we found the residential segregation index between nonwhites resident in the metropolitan area 5 years or more and NWNP to be 80.5; comparing nonwhites with less than 5 years in the metropolitan area and NWNP, the index was 81.0. Comparing the nonwhites who were in the metropolitan area 5 years before with those who were outside 5 years before, the index is 13. Thus the recent nonwhite in-migrants to Chicago are distributed somewhat differently from the rest of the nonwhite population, but are equally segregated from the native whites. It is not possible to interpret these results in terms of the general assimilation and dispersion processes under consideration.

[21]Amos H. Hawley, "Dispersion versus Segregation: Apropos of a Solution of Race Problems," Papers of the Michigan Academy of Science, Arts, and Letters, XXX (1944), 667-674.

TABLE 1.--Selected Characteristics (age-standardized) of Foreign-born and Native-born Ethnic Populations in 1950 and Residential Segregation Indexes of Selected Groups of Foreign Stock from Native Whites of Native Parentage and Negroes in 1930 and 1960; Chicago[a]

	% high school graduates (male)		% income <$3,000		% white collar (male)		Residential segregation vs. NWNP			Residential segregation vs. Negro		
	FBW	NFMP	FBW	NFMP	FBW	NFMP	1930	1960	Change	1930	1960	Change
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
"Old" immigrant groups												
England and Wales	44	50	52	45	50	51	11	18	+7	84	83	-1
Ireland	24	47	57	48	22	47	23	31	+8	85	84	-1
Norway	31	47	50	46	24	50	44	37	-7	90	90	0
Sweden	25	48	46	44	23	50	26	30	+4	90	86	-4
Germany	37	34	51	49	35	42	22	19	-3	88	88	0
"New" immigrant groups												
Austria	29	40	51	46	34	44	30	16	-14	89	88	-1
Czechoslovakia	25	32	59	49	23	36	59	37	-22	93	89	-4
Italy	15	27	56	51	24	36	52	32	-20	79	81	+2
Poland	18	25	62	54	26	30	63	38	-25	94	93	-1
U.S.S.R.	35	60	47	37	60	74	51	44	-7	90	90	0
"Newer" immigrant groups												
Mexico	14	16	65	73	8	13	71	54	-17	77	73	-4
Puerto Rico[b]	13	29	86	72	22	36	NA	67	NA	NA	69	NA

[a]Foreign stock is the foreign-born (FBW) plus the native-born of foreign or mixed parentage (NFMP). NWNP is native white of native parentage. Characteristics refer to the Standard Metropolitan Area population, while segregation indexes refer to the city population. NA means not available.

[b]Socio-economic characteristics for Puerto Rican population refer to U.S. total. Puerto Rican population by Community Areas for Chicago available for 1960 only.

Sources: Characteristics from U.S. Bureau of the Census, U.S. Census of Population: 1950 Vol. IV, Special Reports, Part 3, Chapter A, Nativity and Parentage, and Chapter D, Puerto Ricans in Continental United States. Distributions of population by Community Area for 1960 from the 1960 census tract bulletin for Chicago, and for 1930 from Ernest W. Burgess and Charles Newcomb, eds., Census Data of the City of Chicago, 1930 (Chicago: University of Chicago Press, 1933).

TABLE 2.--Selected Socio-economic Characteristics (unstandardized) of White and Non-whites: City of Chicago, 1940, 1950, and 1960

Characteristic	Non-white	White
Residential segregation index, whites vs. Negroes		
1930		85
1940		85
1950		79
1960		83
% high school grad., ages 25+		
1940	16	25
1950	25	37
1960	29	37
% white collar, male		
1940	17	40
1950	17	41
1960	21	40
% home-owners		
1940	7	26
1950	12	33
1960	16	39
% of multiple-person households with 1.01 or more persons per room		
1940	41	17
1950	46	14
1960	34	10

Source: Data for 1940 from the 1940 census tract bulletin for Chicago; for 1950 from Philip M. Hauser and Evelyn M. Kitagawa, eds., Local Community Fact Book for Chicago, 1950 (Chicago: Chicago Community Inventory, 1953); and for 1960 from the 1960 census tract bulletin for Chicago.

TABLE 3.--Selected Socio-economic Characteristics (unstandardized) of Puerto Ricans and Non-whites: City of Chicago, 1960

Characteristic	Non-white	Puerto Rican
Residential segregation vs. whites	83	67
% high school grads., total	29	11
Median family income	$4,742	$4,161
% of families earning <$3,000	28	27
% of families earning >$10,000	9	4
% of home-owners	16	6
% substandard dwellings	26	33
% 1.01 or more persons per room	34	52
% h.u.'s built since 1940	12	6
Median gross rent	$88	$79
Median number of rooms	3.9	3.7
Median number of persons	3.0	4.0

Source: Data are from the 1960 census tract bulletin for Chicago.

STUDIES OF COMPARATIVE PREVALENCE OF CARDIO-RESPIRATORY DISEASE IN ENGLAND AND THE UNITED STATES

Walter W. Holland, B.Sc., M.B., B.S., Visiting Lecturer
Dept. of Epidemiology, Johns Hopkins School of Hygiene

I would like to begin by saying that this work has been largely stimulated by Professor D. D. Reid and has been done in conjunction with Dr. G. A. Rose of the London School of Hygiene; Dr. T. Mork and Dr. S. Humerfelt of the University of Bergen, Norway; Dr. E. B. Raftery, Dr. R. Seltzer and Mrs. P. McPherson of the Johns Hopkins University, School of Hygiene; Dr. R. Markush of the National Heart Institute and Dr. R. W. Stone of the New York Telephone Company.

Dorn has shown a wide disparity between the United States and Great Britain in mortality from coronary heart and chronic respiratory disease. In view of the many factors which can affect the recording of mortality from particular diseases in different countries it seemed worthwhile to investigate whether these differences reflect a real difference in the prevalence of these diseases in Great Britain and the United States.

At first glance it would appear that one should examine a random sample of the total population of each country to achieve this goal. This is obviously beyond our resources. Surveying the population of one or more areas in the two countries would also not serve our purpose as no one area is representative of the country as a whole. Comparison of hospital admissions or discharges in the two countries suffers from the drawback that criteria of admittance to hospitals differ greatly because of different medical care arrangements.

The intention in our studies has been to select one factor at a time in order to test hypotheses erected on the basis of these mortality studies. In order to compare disease prevalence in different geographic areas we have used occupational groups. We have thus obtained populations which are reasonably uniform as to job, age, social class and sex. The aim of our studies has been to examine all men in a particular area aged 40 years or more, who are doing a certain job, i.e. we have used small "chunk" samples. We have tried to select groups doing similar work in each area. Medical criteria for employment, discharge and disability are also similar. It has been possible to obtain ancillary information such as deaths, retirements and sickness absence in the last three years in the groups surveyed so that some of our comparisons can also be retrospective.

In London we examined Central Post Office van drivers. To compare with these we used Post Office van drivers and men classified as van drivers by the Telephone Company in three English country towns (Peterborough, Norwich and Gloucester). In England the postal and telephone services are both under the same management. In the United States we examined outside telephone workers in three areas -- Washington, Baltimore and Westchester (N.Y.). These are men employed by the Plant Department of the Telephone Company, all of whom are classified as drivers, who do such jobs as construction, installing and repair work. These are all men who mainly work outdoors. The results from the postal and telephone employees, in England, did not differ, and so have been combined. The examinations have all been performed during May and June of 1960, 1961 and 1962. Excellent co-operation was secured from both Management and Union in all areas. Scheduling of the men for examination was the responsibility of Management, while persuasion to attend was the task of the Union.

Each examination lasted about thirty minutes. We used the British Medical Research Council Questionnaire for respiratory disease symptoms. For cardio-vascular disease we used a questionnaire developed by Dr. G. A. Rose, and approved by W.H.O., to elicit symptoms of angina pectoris, possible myocardial infarction and intermittent claudication. Questions on residence and smoking habits were also asked. A 13-lead electrocardiogram was obtained on each individual. Blood pressure was measured in random order both by the standard method and one designed to reduce observer variation. Lung function was determined by use of a simple machine (McKesson Vitalor) which measures the one-second Forced Expiratory Volume. Weight, standing and sitting height, and skinfold thickness (triceps and subscapular sites) were measured, and all these examinations were always performed in same order. Each man on leaving the survey team was given a small container in which he was asked to collect all sputum brought up from his chest in the first hour after arising the next day.

Great care was taken to reduce observer variation. I participated in all three surveys. Various methods were utilized to ensure standardized use of the questionnaires. These included listening to special records of real and simulated interviews designed to illustrate some of the problems. Prospective observers also performed test interviews with trained observers present. All interviews, in any case, were recorded and checks on the observers were made both during and after each survey. A special caliper (developed by Dr. G. A. Rose) was used for placing the chest electrode uniformly in taking the electrocardiogram. Uniform instructions were given in the performance of the lung function test.

Before discussing the results I would like to emphasize that though I may appear to be generalizing to the population as a whole, I am well

aware that this is quite unjustifiable at this stage. I would certainly welcome any comments that you may have on the use of such small "chunk" samples in such studies. I hope I will be forgiven for using such euphemisms as London, Country Towns and the United States. I must also point out that though tests on uniformity have been performed on data from the Country Towns, no such procedures have yet been carried out on the American data, and pooling may not be justifiable. I should also like to say that the results presented are based on preliminary hand tallies and are subject to alteration.

RESULTS:

(1) Respiratory Symptoms

There is very little difference in the frequency of single symptoms such as cough or phlegm between the three areas. Londoners have, however, more shortness of breath than the others.

When, however, symptoms recorded for an individual are combined, differences between the three areas appear. Thus, persistent cough and phlegm with one or more episodes of increased cough and phlegm are almost twice as common in each age group in Londoners as in country dwellers and Americans. An even greater difference is seen in the frequency of individuals who have persistent cough and phlegm as well as dyspnoea grade III or more (this is one of the proposed definitions of chronic bronchitis).

It would thus appear that while the prevalence of the single symptoms of respiratory disease is the same, there is a difference in severity.

(2) Sputum

This difference is also reflected by the sputum returned. There are many more specimens of muco-purulent sputum in the English groups than in the American, and furthermore the distribution of the volume of the samples returned shows a gradient London - Country Towns - United States.

(3) Smoking Habits

Before considering disease manifestations in different geographic areas any further, it is important to look at the smoking habits of these groups. It is seen that the English groups resemble each other closely. There are, however, a greater percentage of non-smokers in the American groups than in the English, but those who do smoke in America, smoke more heavily.

(4) Smoking and Respiratory Symptoms

There is a great difference between smokers and non-smokers in each area in the prevalence of respiratory symptoms. Smokers have more symptoms than non-smokers in each area.

(5) Lung Function

As a measure of lung function we have used the one-second Forced Expiratory Volume, which is one of the simplest and best single measures of bronchial obstruction. This measure of lung function decreases with age in each area; it is lowest in Londoners and highest in the United States group.

As smoking habits differ within each age group we have calculated the regression coefficient of forced expiratory volume on age and have used this to standardize the $F.E.V._{1.0}$ of each individual to age 40. By this means the effect of smoking on the $F.E.V._{1.0}$ is also evident in each area and the differences between areas remain.

(6) Sputum Production and Smoking

The effect of smoking on sputum production is shown in the distribution of volume of sputum. In each area smokers bring up more sputum than non-smokers, but the differences between areas persist after allowance for smoking. Smoking habits appear to have little effect on the type of sputum returned.

(7) Prevalence of Cardiovascular Symptoms

It may be seen that while there is little difference between London and the United States in the prevalence of angina, there is a marked difference in the prevalence of a possible past history of myocardial infarction. The latter is more common in the United States.

(8) Blood Pressure

Mean systolic and diastolic blood pressure is lowest in each age-group in London. This difference between town and country dwellers has also been demonstrated by others. In the 50-59 year age-group, mean blood pressure is higher in the United States group than in country towns. This difference in means is not a result of a few extreme values in one series.

(9) Electrocardiographic Abnormalities

The electrocardiograms which were obtained in the surveys were classified by actual changes rather than by diagnostic categories. The method of classification was devised by Blackburn of Minneapolis and is designed to reduce inter- and intra-observer variation. All English tracings were read by one observer, while all American tracings were read by another. No comparison between these two observers has yet been made. No consistent difference in frequency of abnormalities between the three areas are seen, except that shown by Q and QS changes and T-wave changes both of which are indicative of a major Ischaemic event, such as an infarction. These changes are commoner in the United States group than the English, but until comparisons between the observers have been made no firm conclusions can be drawn.

(10) Anthropometric Measurements

It may be seen that the English are not only shorter and lighter, but also less fat than those surveyed in the United States.

In conclusion, I believe that the methods described are useful in measuring the prevalence of disease in different geographic areas. The differences in mortality from respiratory disease in England between town and country, and between Great Britain and the United States are reflected by these studies when comparisons are based on combinations of marked symptoms. The difference in severity of disease in these areas is also shown by differences in sputum production and lung function. Differences in cardiovascular disease are less clear-cut. While there is a difference in the prevalence of possible past myocardial infarction, no such difference is evident for angina.

II

CONTRIBUTED PAPERS

Chairman, Edwin D. Goldfield, U. S. Bureau of the Census

Page

Statistics from the National Sample of Marriage and Divorce Transcripts - Hugh Carter, Carl E. Ortmeyer and Alexander Plateris, U. S. Department of Health, Education, and Welfare.. 24

Measuring Wage and Employment Changes as Workers Age by Use of Cohort Data - David J. Farber, U. S. Bureau of Old Age and Survivors Insurance..... 34

Sample Size Determination in Simple Analytical Surveys (Abstract) - Joseph Sedransk, Harvard University..... 45

The Life Table from the Standpoint of Statistical Analysis (Abstract) - Norman B. Ryder, University of Wisconsin..... 45

A Comparison on the Basis of the Nashville Morbidity Survey of Two Frames with Equal and Unequal Sized First-Stage Units - T. Khosla, University of Aberdeen, A. L. Finkner, Research Triangle Institute and J. C. Koop, North Carolina State College..... 46

STATISTICS FROM THE NATIONAL SAMPLE OF MARRIAGE AND DIVORCE TRANSCRIPTS

Hugh Carter, Carl E. Ortmeyer, and Alexander Plateris
U. S. Department of Health, Education, and Welfare, Public Health Service
National Center for Health Statistics, National Vital Statistics Division

For the calendar year 1960 a national sample of marriage and divorce transcripts was collected for the United States. Prior to 1960, and for many years, national statistics were prepared from tables compiled in State vital statistics offices. After checking for internal consistency and adding estimates for missing areas, reports were issued giving information on the number and characteristics of persons at marriage or divorce.

The limited value of such data need not be argued here. While many States had excellent vital statistics organizations, others were seriously understaffed, and frequently there were no resources for essential field work with local officials. From the standpoint of national statistics the most serious deficiency--surprising to many--is the fact that 7 States still lack central files of marriage records and 11 do not maintain State files of divorce records.

The United States is an "underdeveloped area," in comparison to many Western nations, so far as statistics of marriages and divorces are concerned. While reports of these events in the United States go back over 80 years, only widely spaced special surveys are available for early decades, and the data are subject to the usual deficiencies of such procedures. Beginning with 1922 there are excellent annual reports of marriages and divorces, based on mail questionnaires, but these end with 1932; data for the following decade are fragmentary. The pretabulated figures, begun in 1948, have improved steadily, but at their best leave much to be desired.

As part of the programs to improve national statistics of family formation and dissolution, reporting areas were established to which States were admitted upon meeting specified standards. The Marriage Registration Area now includes 35 States, and the Divorce Registration Area includes 21 States.

For the year 1960, with a view to obtaining comprehensive national data, a sample of records was collected from all 50 States. There is here reviewed the sample design, the collection process, the data obtained, and a brief evaluation.

The Sample Design

Relatively small probability samples of marriage and divorce transcripts were selected for 1960. As a result, some detailed statistics for local areas previously reported on a pretabulated basis will not be published. The overall sizes of samples to be secured, based on 1958 returns, were expected to be about 41,000 marriage records and 16,000 divorce and annulment records from annual totals of approximately 1.5 million marriages and 370,000 divorces. The objective of the sample was to secure an array of statistical estimates using information from relatively small numbers of sample records. Annual and monthly totals were to be estimated from the samples and compared for consistency with corresponding annual totals based upon pretabulated reports from the States. The samples themselves were intended to provide estimates of the distributions of several characteristics, such as age, number of previous marriages, color, and residence. Marriages and divorces by county of occurrence were reported on a pretabulated basis. Estimates were made for each of four area units: (a) each State in a Registration Area; (b) the entire Registration Area; (c) each region; and (d) the United States. Each of these estimates was to fall within specified maximum sampling error tolerances of the characteristics that would have resulted from a complete census of all the records. Several of these maximums are shown in Table 1.

Apart from the small over-all sizes of the samples, two other factors loomed as important in relation to the design of the sample, especially for States not in the Registration Areas: local, county, and court officials with little or no experience in sample selection would be expected to select the samples and prepare the transcripts; State and Federal officials would have to move quickly in establishing contacts and securing cooperation from local officials who, if they had been participating in State-Federal vital statistics programs in any way, had been doing so to a much more limited extent. Several problems appeared later in securing the requested sample records from local officials without previous experience in providing records for vital statistics uses.

The samples of marriage and divorce records are similar in design. In each sample, the reporting areas, i.e., States with suitable central files of records and local areas in other States, were divided into six strata, the characteristics of which are shown in Table 2. Strata 1-4 consist of States in the Registration Area, divided according to expected totals of records. With respect to estimates for each of these States, it was determined that a sample with a minimum size of 400 records would satisfy maximum sampling error specifications. To simplify sampling procedures, 4 sampling rates were used for these State samples--1, 5, 10, and 100 percent in the marriage sample, and 5, 10, 50, and 100 percent in the divorce sample. Thus, the design for the Registration Areas represented a reconciling of the objectives of holding samples close to minimum sizes but designating only a few readily comprehended sampling rates. Specifying a minimum for the number of records required in each State, along with use of only 4 sampling rates, plus the fact that 1960 totals were, in most cases, larger than 1958 totals, resulted in sample returns

large enough to provide estimates with errors falling well within the initially specified maximums.

The portion of each sample from States not participating in the Registration Area was selected independently in each of the 4 U. S. regions. Selection was carried out in two stages. In the case of marriages, the first stage sampling units consisted of 8 States and the District of Columbia, which were sampled from central files of records, and all counties in 9 States and New York City without central files complete enough to be used for sampling. For divorces, the first stage sampling units consisted of 3 States and the District of Columbia sampled from central files and all counties in 29 States sampled from local files.

First stage sampling units were selected systematically with probabilities proportionate to size. The measure of size for each such unit (or reporting area) was the total of records reported for 1958. In each region, these sample units were then sorted into strata 5 and 6.

Stratum 5 included all States sampled from central files, along with those counties with expected total records equal to or exceeding the first stage skip interval. All these areas were selected with certainty. In stratum 5, individual records were subsampled systematically with a skip interval one-fifth as large as the first stage skip interval. Since all areas in this stratum were selected for the sample with certainty, the over-all sampling rate for each record equals the rate for the second stage, i.e., 5/X where X is the skip interval applied in the first stage.

The remaining counties sampled in the first stage constituted stratum 6. In stratum 6, an expected subsample of 10 records was to be selected from each sample county. Hence, the expected total (1958) for each area was divided by 10, and the nearest whole number to this quotient was the skip interval for subsampling of records in 1960. In this stratum, the over-all probability of any record being selected in the sample was expected to be $(N_h/X)\ (10/N_h)=10/X$, where N_h is the expected total of records for the sample county and X is the skip interval used in selecting the county. This over-all probability is uniform throughout the stratum regardless of variations in expected totals among the sample counties, or of variations for any county between its expected (1958) total and its reported 1960 total. If, for example, the expected total of the h'th county was N_h, but the total reported for 1960 was $2N_h$, the over-all probability of selecting any record then becomes $(N_h/X)(20/2N_h)=20/2X=10/X$.

It should also be noted that the over-all sampling rate in stratum 6, 10/X, is double the over-all for stratum 5, 5/X. The higher rate in stratum 6 was necessitated by the fact that the sampling error arose both from the sampling of areas and the subsampling of records, whereas in stratum 5 the areas were selected with certainty and contributed nothing to the sampling error. Thus, the higher sampling rate in stratum 6 offsets the greater sampling error.

Data Collection

In order to obtain the information necessary for planning the marriage and divorce transcript program, questionnaires went to State offices of vital statistics maintaining central files of marriage or divorce records for information about registration procedures, preferences concerning reporting schedules, type of images to be provided, and payment rates. Purchase agreements were concluded with 41 States in the marriage program and 21 in the divorce program. In the remaining States, agreements were concluded with officials of the counties falling in the sample. In many counties more than one court is empowered to grant divorce decrees, and the cooperation of the clerks of all such courts was necessary.

Almost all local areas agreed to participate in the programs, though some officials were unwilling to select the actual samples. In such cases, the sample was drawn by State or Federal officials, or by social scientists [1] or persons recommended by them. [2] Thus, all selected counties, except seven, participated in the sample: one county in the marriage program and six counties in the divorce program did not participate.

Eventually, the sample transcripts totaled about 42,000 marriages and 17,000 divorces; the Registration Areas provided about 39,500 and 13,500 transcripts, respectively. The transcripts sent by the States were copies of the original records. Sample cases from counties were reported on specially designed short forms, which included only a few basic items of information.

Late reporting, incomplete or incorrect sample selection, and illegible microfilm images were the main reporting problems; an extensive query program conducted by the NVSD made it possible to adjust most of these. The number of sample transcripts received from the State offices were complete except for one State in the marriage program and two in the divorce program, where unsatisfactory registration procedures made it impractical to select sample transcripts covering all the events that occurred in 1960. For the county samples, in addition to the seven refusal counties, a few other local areas sent in incomplete samples. Altogether, about 0.6 percent of all marriages and 1.8 percent of divorces were not covered by the sample. Data on these cases will be published as unknown.

Most of the transcripts were coded and

punched in the NVSD; some of this work was done under contract by another agency. The marriage sample is being processed on electronic computer units and the divorce sample on conventional punched card equipment.

Data Obtained from the Sample

Data to be published from the 1960 sample of marriages are more limited for the 4 regions and for the United States as a whole than for the Marriage Registration Area, but a few basic items will be available. Where appropriate, separate data will be published for brides and for grooms. In addition to counts of marriages, these are:

- County and State of occurrence of the marriage
- Date of marriage, from which distributions by month and by day of week are available
- Age
- Race or color
- Whether the marriage is a first or remarriage, along with marital status prior to the marriage

The securing of this information represents an advance for the program since, for a number of nonregistration States, the only recent data have been marriages by State and month of occurrence. Information on age at marriage is 99 percent complete for the United States.

Four other variables will appear in the annual tables published for the Marriage Registration Area:

- Auspices of the ceremony, that is, civil vs. religious
- Marriage order, i.e., 1st, 2nd, 3rd, etc.
- State of residence (or foreign country)
- County of residence

Among these items, State and county of residence are the most completely reported. It will be possible to present more detailed data on residential propinquity of couples at marriage, as well as on proportions of marriages which occur in counties of residence of brides and grooms. With the completeness of reporting in the MRA States of county of residence and data on age and on marriage order, the way is opened for analyses of the interdependence of variation in age-adjusted marriage rates at first marriage and at remarriage with variables believed to affect these rates, such as income levels, degree of urbanization, rate of population growth, educational levels, proportion of the labor force in manufacturing, and rates of unemployment. While marriage rates can not be computed for individual counties, all counties in the MRA can be sorted into a few categories on any one or two of the socioeconomic variables, and comparisons can be made of marriage rate variations among any set of such categories.

Finally, in addition to items available for the entire nation and others available for the MRA, several are available from small numbers of States within the MRA. Six States report the church denomination of clergymen officiating at religious ceremonies, and two report this item for brides and grooms; two report grades of school completed; fifteen report occupation and industry (brevity of the information recorded in many cases made this item difficult to process); one State reports occupation of the fathers of the marital partners; and two report race or color of the parents. In general, items available only from limited numbers of States will be tabulated and published, if completely enough reported, in special studies.

Divorce totals for 1960--national, State, and county--are more reliable than those for past years. This was made possible by the additional information on registration procedures obtained during the data collection, and by the comparison of figures reported by State officials with estimates based on sample transcripts.

The sample program produced detailed national and regional divorce statistics for duration of marriage, children of the divorced couple, month of marriage, and the area where the decree was granted and where the marriage was performed.

For the DRA and each DRA State some adtional data were obtained, and where appropriate, will be published separately for husbands and wives. Those reported completely enough for analysis are: legal grounds for decree, plaintiff, person to whom the decree was granted. Data for the total DRA and for each registration State will be published for the first time for 1960.

Unsatisfactory reporting and sampling variability made it impossible to tabulate national, regional, and DRA statistics with the amount of detail which was used in the past for a limited number of States. However, for a few States various detailed crosstabulations of personal characteristics, such as age, marriage order, race or color, and area of residence, can be prepared. Individual States reported some previously unavailable data, such as information on duration of marriage to separation and duration of separation to decree, decisions on custody of children, and number of children of the marriage being dissolved.

Evaluation

There are three aspects to our evaluation of this first national sample of marriage and divorce transcripts: (1) sampling error estimates and information about nonsampling errors; (2) needed steps to reduce these errors; (3) important gains registered by the 1960 project and future implications for improved annual data.

1. Sampling and nonsampling errors.

Turning to sampling errors, several preliminary estimates have been made, but not yet published. These include the sampling errors for specified percents of the totals for the MRA and DRA and for each Registration Area State.

In general, these sampling errors are markedly less than the maximum sampling errors shown in Table 1. Sample returns for individual Registration Area States were considerably larger than the specified minimums of 400 transcripts. Among 33 MRA State samples only one had fewer than 500 records and six consisted of more than 1,000 records each. These larger returns reduced the sampling errors of estimates for most of the Registration Area States and for the entire areas well below the original maximums. They will also permit the tabulation of somewhat more detailed age and other distributions, particularly for each Registration Area as a whole.

Sampling errors for median and mean age estimates, as well as all sampling error estimates for the four regions and the United States, will be computed and published in detail.

Through administrative procedures followed in selecting, inspecting, and processing data for both 1960 programs, four sources of nonsampling deviations or "errors" were identified. The first of these arose because of difficulties in segregating the population of events to be sampled. This was especially marked in sampling divorce records. In one State the divorce records for the first three months of the year 1960 could not be sampled, given the resources available for this part of the work; hence, the sample was drawn from decrees granted from March 1, 1960, up to March 1, 1961. More common than this unique type of difficulty were cases in which some records of divorce decrees granted or of marriages performed during 1960 were filed among records of a year prior to 1960. In a few States the marriage license is good indefinitely. Divorce decrees were sometimes filed by date of the original complaint, hence, any decree granted in 1960 in a case continued from a preceding year would appear in the files for that year. Complexities in the divorce laws of some States led to difficulties in determining exactly when a decree became final and one or both parties were free to remarry. It was not possible in the case of marriages to separate so-called second ceremonies (usually a religious ceremony following a civil ceremony for the same couple) from other records if separate licenses were issued for both.

The second source of nonsampling errors was in selecting the samples of records. Controls on such mistakes were fairly complete for Registration Area States. In all areas, if these mistakes were not caught earlier, they were usually identified when inquiries were made about discrepancies between reported annual totals and number of sample records submitted. This activity, as well as queries about missing numbers on the records, consumed a large amount of staff time, but it yielded much of our information about nonsampling errors other than incompleteness of item reporting.

The third source of error was mistakes made in coding and punching the data. Detailed reports from the unit which coded and punched the 1960 marriage records indicate the low error ratios in coding of between 0.1 and 0.3 percent. These ratios are the proportions of all items which were coded erroneously. Errors made in coding the divorce records were of approximately the same magnitude. Errors made in punching were negligible. Since this was a new program, both coding and punching were verified 100 percent.

The fourth, and the major source of nonsampling error arose from the incompleteness with which various items of information were reported.

In Tables 3 and 4 are set forth preliminary data on completeness of reporting of marriages and divorces showing percents of items incomplete. 3/

The following conclusions emerge from the marriage data:

(1) Ages of brides and grooms are reported with great completeness in records from all areas--thus, data become available for age-specific and age-adjusted marriage rates.

(2) County and State of residence of both brides and grooms are reported with exceptional completeness in the MRA States--thus, studies of residential propinquity and computation of marriage rates by areas of residence at marriage become feasible.

Other conclusions based on more detailed, unpublished data are:

(1) The distinction of civil vs. religious marriages in the MRA was reported with 90 percent completeness.

(2) Place of occurrence and date of marriage are reported with uniform completeness, except for a 2 to 3 percent loss in the Southern Region.

(3) Records from the Northeast Region are most complete; the greatest degree of incompleteness appears in the Southern Region.

(4) Over 80 percent of the loss of information about race or color of brides and grooms resulted from the absence of an item requesting this information on the record forms of a few States and counties.

(5) Roughly two-thirds of the loss of data

about previous marital status and marriage order resulted from the absence of the item on the forms of several reporting areas; marriage order is much more complete for the MRA than for the non-MRA.

To summarize the data on incompleteness of marriage reporting, the largest source was the absence, on the marriage record forms of several reporting areas, of the items of information desired, especially race or color, previous marital status, and number of the marriage. The next largest source was the failure to complete items of data on forms on which such items were requested. Finally, least in importance was a very small number of sample records which were not received.

Under-reporting of items on divorce records is the major deficiency of the program. In 1960, the completeness of reporting for the Divorce Registration Area of the age of husband and of wife was only 41 percent, varying among the States from 100 percent to 4 percent. Percentages of completeness of other characteristics of the spouses were similarly low. In order to improve reporting of these variables, a major effort by Federal and State agencies and others will be necessary.

The 1960 data collection process was not designed to make comprehensive tests of the reliability of items reported on the records. Special studies of this source of nonsampling error are now being planned.

2. Steps to reduce errors.

The steps that can be taken to reduce the sampling errors, and thereby to make possible the presentation of more detailed statistical series, are fairly simple to specify, when compared to recommending steps for reducing the nonsampling errors, particularly the incompleteness of reporting.

An approximate doubling of the marriage sample is in prospect for the 1963 calendar year. Another step which will reduce the sampling errors for marriage estimates is to establish sampling rates for each MRA reporting area such that the sample returns from all areas are more nearly equal than was the case in 1960. This same device should also reduce the DRA State sampling errors. Plans are also being formulated for again securing samples of marriage records in the near future from the few States still not in the MRA. This should stimulate interest in these States in improving the completeness of their reporting. First priority in the divorce statistics program will be given to reducing the incompleteness of reporting and other sources of nonsampling errors.

There are several specific technical steps and, more important, broader programs for building improved marriage and divorce reporting which can be taken as rapidly as available staff and resources will permit. Among the technical points are:

(1) Fullest use of techniques for checking totals of records reported to States and to the NVSD. (It is helpful to secure independent counts of fees paid for marriage licenses or of premarital blood tests, or court reports to judicial agencies of totals of divorces and annulments.)

(2) Payment arrangements at State and Federal levels which reward those areas promptly reporting complete data.

(3) Clear assignment of responsibility at the local level for securing and reporting all information on the records.

(4) Filing or indexing of the records at local levels by date of marriage ceremony or of divorce decree.

(5) Provisions for prompt return of marriage records to issuing clerks, and for prompt reporting of both marriage and divorce records from local to State officials.

(6) Regular programs of querying suitable sources of information about items missing on the records when the records are inspected by local clerks and State registration personnel.

Broad programs for encouraging interest in marriage and divorce registration of high quality are under way and will probably have great impact on the quality of the data. Arrangements are now being carried out to hold conferences with local registration officials in States where this approach shows promise of success. Certain studies of interviewer effectiveness in survey research have found that high interest in the results is a key characteristic of the best interviewers. Thus, it seems desirable to demonstrate to the local officials the uses of marriage and divorce statistics. Emphasis will also be placed on encouraging questions about programs and procedures.

Efforts are being made to interest more university and State research units in the use of these records. Adequate budgeting to compensate registration officials for their services in making the records available is being recommended. The major professional associations concerned with registration and use of these data have endorsed the registration area approach to improving their quality.

3. Gains from the 1960 programs.

Much remains to be done in providing the United States with adequate systems of marriage and divorce registration which can produce

complete and accurate statistics at least on the most essential items. However, certain important gains have been made. Counts of the records have been improved in many areas. Requests for the sample records, along with queries concerning missing information, inconsistencies in counts of records, and other problems have supported efforts of registration leaders in several States to secure improvements. It is reported that visits by State and regional representatives to local officials in arranging for selection of the samples have heightened interest in other vital records.

Specific gains in amounts and types of data available have also resulted. Data on age at marriage of comparable completeness have never before been available for the entire nation. Counts of divorces and annulments, although leaving much to be desired, probably are more accurate than at any time in the past. Data on counties of residence of brides and grooms for the entire MRA represent a marked gain. While demographic items, such as age at divorce and race or color, are completely reported in only four States, a number of legal items of interest to students of legal procedures involved in the dissolution of marriage are available even for non-DRA areas. Examples are the distribution of divorces by legal grounds, the extent to which husbands or wives are plaintiffs and defendants, or are granted decrees, and duration of the marriages being terminated.

The 1960 sampling operation has not only produced better data; it has also increased interest in the program. Social scientists who helped us frequently expressed surprise at the registration conditions existing in their own States. Public officials also learned more about the program and indicated a desire to improve registration. As a logical consequence, we expect an accelerated growth of the Registration Areas until truly national statistics are made possible.

FOOTNOTES

1/ We are grateful for the assistance of the following social scientists in collecting sample transcripts in counties where these were not otherwise available, and in recommending other individuals qualified to do so:

C. Wylie Alford
Sarah F. Anders
William W. Biddle
Melvin S. Brooks
William Bruce Cameron
Paul J. Campisi
Alice S. Christensen
Harold T. Christensen
James E. DeBurger
Otis Durant Duncan
Everett D. Dyer
Thomas Ford
Philip M. Hauser
Morton B. King, Jr.
Clifford Kirkpatrick
Clyde V. Kiser
Wilfred G. Marston
Floyd Allen Pollock
Austin L. Porterfield
Douglass B. Radabaugh
Calvin F. Schmid
Virgil L. Seymour
Gordon Shaw
Leonard M. Sizer
Martin Taitel
Donald L. Taylor
William T. Tucker
James D. Turner
Joseph C. Urbon
Walter T. Watson
Raymond H. Wheeler
Alvan O. Zarate

2/ The following persons with specialized knowledge of local conditions were of invaluable assistance to us in collecting sample records from local files:

James W. Adams
Betty Bowling
Rowena Colling
Gladys Combs
Margery Darnell
Shirley J. Davis
Mrs. E. P. Harned
Mrs. Clyde Hopkins
Mary Ann Leeper
Kay Mabry
Clark McWhorter
Marilee Morgan
Vernon Son

3/ Only data for the United States and each Registration Area are shown. Estimates for each region and each Registration Area State are also available.

Table 1.--MAXIMUM SAMPLING ERROR TOLERANCES (2σ) FOR SPECIFIED ESTIMATES FROM SAMPLES OF 1960 MARRIAGE AND DIVORCE TRANSCRIPTS

Type of Record and Area	Percent of Cases with Characteristic				
	2	5	10	25	50
Registration Area State (Marriage or Divorce)	1.4	2.2	3.0	4.3	5.0
Marriages:					
Registration Area (MRA)	0.3	0.6	0.7	1.0	1.1
Each region	1.5	2.4	3.3	4.8	5.7
United States	0.8	1.2	1.5	2.2	2.7
Divorces:					
Registration Area (DRA)	0.4	0.4	0.8	1.1	1.3
Each region	2.0	3.0	5.7	6.0	6.9
United States	0.9	1.2	1.6	2.4	2.8

Table 2.--CHARACTERISTICS OF THE DESIGN OF THE 1960 STRATIFIED MARRIAGE AND DIVORCE SAMPLE

(Expected figures based on 1958 data)

Type of Area and Stratum	Marriages		Divorces	
	Sampling rate	Expected totals	Sampling rate	Expected totals
United States	--	1,459,000 1/	--	370,000 1/
Registration Area	--	835,000 1/	--	83,000 1/
1	All records	14,000	All records	1,000
2	1/10	43,000	1/2	10,000
3	1/20	393,000	1/10	39,000
4	1/100	385,000	1/20	33,000
Central File States and Certainty Counties	--	434,000	--	192,000
Northeast	1/500	102,000	1/40	19,000
North Central	1/500	146,000	1/80	71,000
South	1/800	103,000	1/160	39,000
West	1/300	83,000	1/80	63,000
Noncertainty Counties	--	190,000	--	95,000
Northeast	--	--	1/20	6,000
North Central	1/221	36,000	1/45	15,000
South	1/374	123,000	1/78	61,000
West	1/173	31,000	1/45	13,000

1/ Includes data for Alaska and Hawaii.

Table 3. INCOMPLETENESS OF REPORTING OF MARRIAGE DATA: UNITED STATES AND MARRIAGE REGISTRATION AREA, 1960

(Numbers not available are weighted totals; each percent is a ratio of number incomplete to corresponding total marriages)

NOTE: Figures are provisional and subject to slight changes.

Variables Not Reported on Sample Records	Area and Estimate					
	United States			Marriage Registration Area [1]		
	Number	Percent	Total Percent Not Available [2]	Number	Percent	Total Percent Not Available
Total Marriages	1,523,381	100.0	100.0	873,224	100.0	100.0
Records Outstanding	9,130	0.6	--	3,480	0.4	--
County of Marriage	3,953	0.3	0.9	203	0.0	0.4
Date of Marriage						
Day	16,363	1.1	1.7	1,033	0.1	0.5
Month	6,780	0.5	1.1	0	0.0	0.4
Age						
Groom	7,242	0.5	1.1	642	0.1	0.5
Bride	7,186	0.5	1.1	986	0.1	0.5
Color [3]						
Groom	227,798	15.0	15.6	74,248	8.5	8.9
Bride	229,432	15.1	15.7	77,182	8.8	9.2
Marriage Order [4]						
Groom	166,090	10.9	11.5	388	0.0	0.4
Bride	163,377	10.7	11.3	398	0.1	0.5
Previous Marital Status [5]						
Groom	200,191	13.1	13.7	42,391	4.9	5.3
Bride	199,122	13.1	13.7	44,042	5.1	5.5

1/ Data on civil vs. religious ceremonies (11.2 percent incomplete), number of the marriage being solemnized (4 percent incomplete), and State and county of residence (less than 1 percent incomplete) are also being tabulated for the MRA.

2/ Outstanding records and variables not reported on sample records, combined.

3/ Approximately 85 percent of the loss of data on color for both the United States and the MRA resulted from the absence of this item on several record forms.

4/ Classification of first vs. remarriage. Approximately 66 percent of the loss of this item for the United States resulted from the absence of necessary items on record forms.

5/ Approximately 70 percent of the data loss on this item for both the United States and the MRA resulted from the absence of the item from several State forms.

Table 4.--INCOMPLETENESS OF REPORTING OF DIVORCE DATA: UNITED STATES AND THE DIVORCE REGISTRATION AREA: 1960

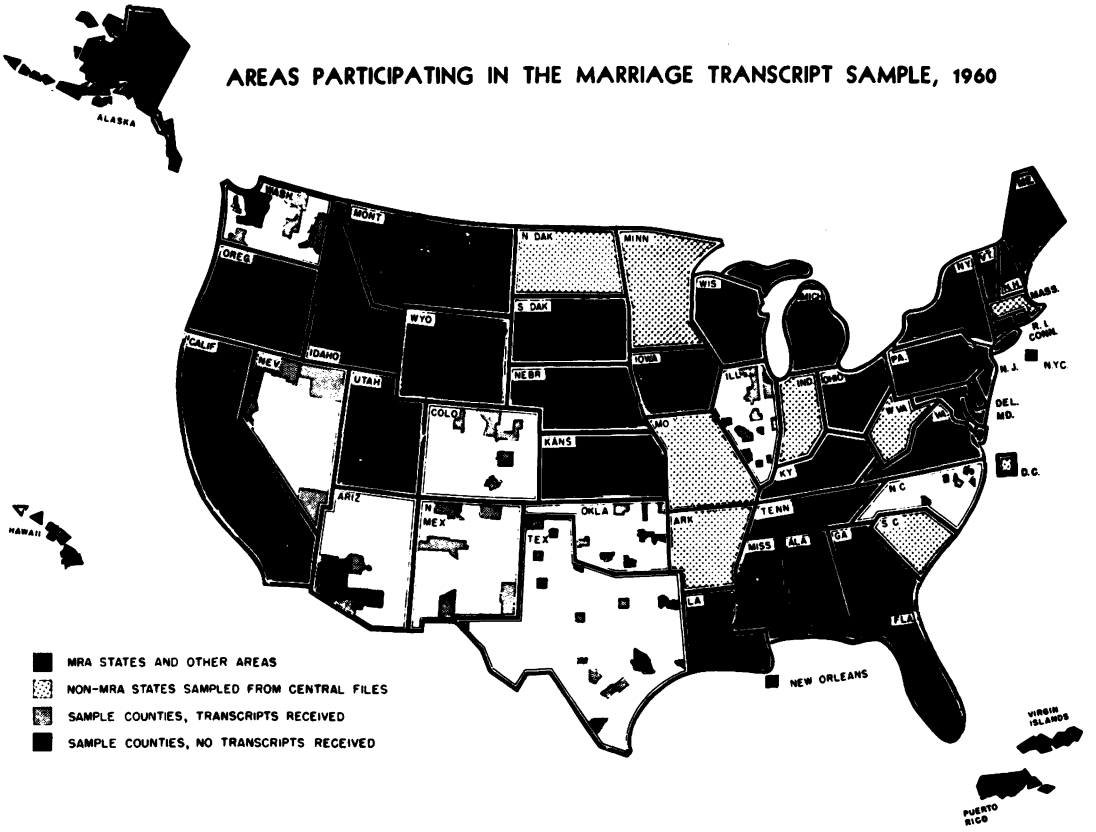
(Figures are provisional and subject to slight changes)

Variables Not Reported on Sample Records	Area and Estimate					
	United States			Divorce Registration Area [1]		
	Number	Percent	Total Percent Not Available [2]	Number	Percent	Total Percent Not Available [2]
Total Divorces	393,814	100.0	100.0	94,074	100.0	100.0
Records Outstanding	6,946	1.8	--	706	0.8	--
Month of Decree	1,510	0.4	2.1	20	0.0	0.8
Year of Marriage	7,050	1.8	3.6	3,430	3.6	4.4
Place of Marriage	53,108	13.5	15.2	17,688	18.8	19.5
Number of Children	18,202	4.6	6.4	11,922	12.7	13.4
Age						
Husband	210,914	53.6	55.3	37,914	40.3	41.0
Wife	208,854	53.0	54.8	37,554	39.9	40.6
Color						
Husband	155,537	39.5	41.3	28,577	30.4	31.1
Wife	152,428	38.7	40.5	27,568	29.3	30.1
Number of Marriage						
Husband	209,285	53.1	54.9	45,345	48.2	48.9
Wife	203,702	51.7	53.5	44,182	47.0	47.7

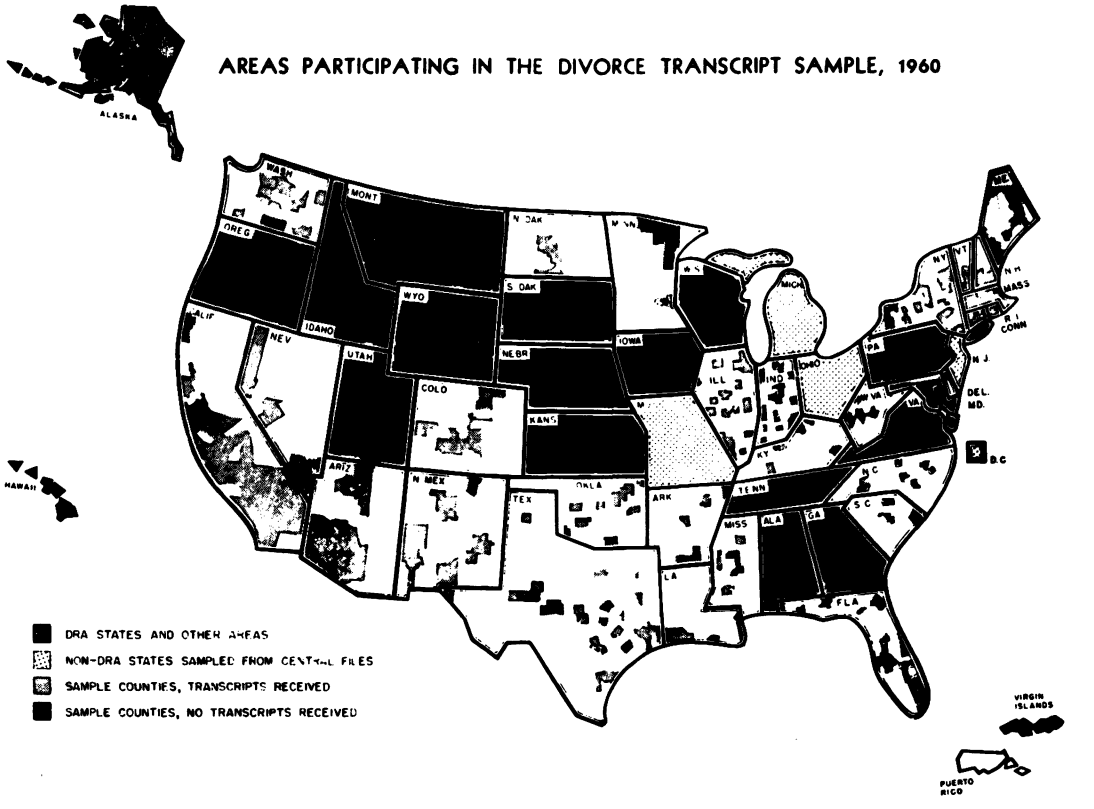
1/ The following other variables are available for the Divorce Registration Area: type of decree (1.8 percent incomplete), plaintiff (2.0 percent incomplete), person to whom decree was granted (4.9 percent incomplete), legal ground for decree (5.2 percent incomplete), county of residence of husband (27.9 percent incomplete), of wife (25.2 percent incomplete).

2/ Outstanding records and variables not reported on sample records, combined.

AREAS PARTICIPATING IN THE MARRIAGE TRANSCRIPT SAMPLE, 1960



AREAS PARTICIPATING IN THE DIVORCE TRANSCRIPT SAMPLE, 1960



MEASURING WAGE AND EMPLOYMENT CHANGES AS WORKERS AGE BY USE OF COHORT DATA

David J. Farber, BOASI

The unavailability of wage statistics for worker cohorts has generally resulted in the use of cross-sectional data to measure changes in wages. Many years ago, Woytinsky cautioned against such comparisons, warning that they

> are likely to be misleading in comparisons over time because of a change in the composition of the labor force; they may show, for instance, a rise of wages in depression because of a reduction in the number of less skilled and low-paid workers, whom it is customary to lay off first ... 1/

Cross-sectional comparisons of average wages from one year to the next, which are based on the wages of employed workers are misleading because they

> do not make allowance for the fact and extent of unemployment ... Based on payrolls, as such an average is, it necessarily measures only the earnings of those who continue to be attached to a job and ignores those who do not have a position and whose names consequently do not appear on the records of any employer. Yet unemployment of course affects the relative well being of those who labor, and its omission results in an overstatement of the actual average amounts received by those who seek work and is an imperfect portrayal of the relative changes from year to year of this larger class ... There is need, therefore, for measuring the average annual earnings of the active members of the working class by thus taking account of the losses of income resulting from the failure to find employment.
>
> The theoretically best method of accomplishing this would be to obtain reports on total annual earnings or income from a large number of identical workmen over a period of years. At present, however, such material is unavailable ... 2/

1/ Encyclopedia of Social Sciences, p. 303. See also Staff Report on Employment, Growth, and Price Levels, Joint Economic Committee, U.S. Congress, December 24, 1959, which at page 142, on the basis of cross-sectional comparisons concludes that from 1947 to 1958 wages tended "to continue to move upward, even during periods of substantial unemployment."

2/ Paul H. Douglas, Real Wages in the United States, Houghton, Mifflin Company, New York, 1930, pp. 11 and 12.

The Cohort and its Selection (Tables 1 and 2)

While Douglas' goal still cannot be fully achieved, it is now possible to trace and analyze the employment and wage histories of a cohort of identical workers as they aged from 1937 to 1957. Neither the employment nor the wage histories of these workers are complete, and they may not be completely representative of the history of the "typical" wage earner. From the wage records maintained by the Bureau of Old-Age and Survivors Insurance, however, it is possible to examine the history of a cohort of 61,202 wage workers--39,018 men and 22,184 women--with respect to that part of their employment covered by the BOASI program and to the wage credits they received as a result of such employment. This paper reports a few of the highlights from a forthcoming study of this cohort which was selected from the Continuous Work History Sample maintained by the Bureau of Old-Age and Survivors Insurance. The cohort represents a 0.1 percent sampling of wage earners working in covered employment in 1957. 3/ Workers were included in the cohort on the basis of two criteria: (1) In 1957, they were working in employment covered by the OASI program; and (2) when working in covered employment in any of the years in the 1951-57 period, they received wage credits 4/ solely from work as wage earners. Selection of the cohort on this retrospective basis guaranteed that the cohort members were alive in the period preceding 1957, and obviated the need for adjusting the wage data to take into account the incidence of mortality. The method of cohort selection, therefore, differs from the more usual method, and conceivably could affect the findings of this study. Our findings with respect to wage differentials and sex differentails, however, are consistent with the findings of other students of wage problems, and suggest that use of this method of cohort selection does not result in findings which would be appreciably different if a different method of selection had been used.

Employment and wage credits of the cohort are analyzed for the 1937-57 period, which is divided into two time spans--1937-50 and 1951-57. Wage histories of workers included in this study

3/ For a description of the sample, see The Continuous Work History Sample Under Old-Age and Survivors Insurance in the United States of America, by B. J. Mandel, First International Conference of Social Security Actuaries and Statisticians, Brussels, November 1956. Also see Jacob Perlman and Benjamin Mandel, "Sampling the Federal OASI Records," Journal of the American Statistical Association, September 1953.

4/ Wages credited to a worker for work in covered employment in a given year for purposes of benefit computations--up to $3,000 for 1937-50, $3,600 for 1951-54, and $4,200 for 1955-57.

Table 1.--Number of Wage Earners in the Cohorts, by Sex, Age, and Average Wage Credits Per Year in Covered Employment a/

Age in 1957	Male					Female				
	Average Wage Credits PYE, 1951-57					Average Wage Credits PYE, 1951-57				
	Total	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Total	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over
Total	39,018	8,989	7,497	9,811	12,721	22,184	10,481	7,215	3,542	946
1-15	324	320	3	1	---	150	148	2	---	---
16-19	3,243	2,812	374	55	2	2,290	1,991	289	10	---
20-24	5,096	2,472	1,882	716	26	3,040	1,658	1,194	188	---
25-29	4,829	870	1,495	1,861	503	2,371	984	902	455	30
30-34	4,775	461	769	1,634	1,911	2,397	1,090	751	472	84
35-39	4,599	419	614	1,267	2,299	2,500	1,064	805	512	119
40-44	3,957	339	486	1,017	2,115	2,390	953	829	463	145
45-49	3,542	324	493	844	1,881	2,326	834	824	502	166
50-54	2,887	275	358	694	1,560	1,788	630	622	382	154
55-59	2,318	216	332	651	1,119	1,365	457	471	296	141
60-64	1,731	148	280	465	838	848	325	293	158	72
65-69	1,064	154	231	334	345	479	216	157	79	27
70 and over .	629	174	168	169	118	234	128	73	25	8

a/ 0.1 percent sample of workers with credits solely from work as wage earners in 1957 and when employed any time in the 1951-57 period.

Table 2.--Percentage Distribution of Male and Female Age Cohorts by Average Wage Credits Per Year in Covered Employment

Age in 1957	Male				Female			
	Average Wage Credits PYE, 1951-57				Average Wage Credits PYE, 1951-57			
	Under $1,200	Under $2,400	Under $3,600	$3,600 and over	Under $1,200	Under $2,400	Under $3,600	$3,600 and over
Total	23.0	42.2	67.3	32.6	47.2	79.7	95.7	4.3
1-15	98.8	99.7	100.0	---	98.7	100.0	100.0	---
16-19	86.7	98.2	99.9	0.1	86.9	99.5	99.9	---
20-24	48.5	85.4	99.5	0.5	54.5	93.8	100.0	---
25-29	18.0	49.0	89.6	10.4	41.5	79.5	98.7	1.3
30-34	9.7	25.8	60.0	40.0	45.5	76.8	96.5	3.5
35-39	9.1	22.5	50.0	50.0	42.6	74.8	95.3	4.8
40-44	8.6	20.9	46.6	53.4	39.9	76.6	94.0	6.1
45-49	9.1	23.0	46.8	53.1	35.9	71.3	92.9	7.1
50-54	9.5	21.9	45.9	54.0	35.2	70.0	91.4	8.6
55-59	9.3	23.6	51.7	48.3	33.5	68.0	89.7	10.3
60-64	8.5	24.7	51.6	48.4	38.3	72.9	91.5	8.5
65-69	14.5	36.2	67.6	32.4	45.1	77.9	94.4	5.6
70 and over .	27.7	54.4	81.3	18.8	54.7	85.9	96.6	3.4

are classified by age in 1957, and by sex. In addition, they are grouped into four subcohorts. Those whose average wage credits per year employed, 1951-57, were: (1) less than $1,200 constitute the low-paid cohort; (2) $1,200-$2,399 constitute the intermediate low-paid cohort; (3) $2,400-$3,599 constitute the intermediate high-paid cohort; (4) $3,600 or more constitute the high-paid cohort.

About two-thirds of the male cohort, and more than nine-tenths of the female cohort, had average wage credits of less than $3,600 per year employed during the 1951-57 period. While this proportion varied with age, among male workers 40 years of age or older, the proportion of lower-paid workers was surprisingly high. For men aged 50-54, 46 percent had average wage credits per year in covered employment (PYE), 1951-57, of less than $3,600, and for those aged 60-64, a little more than one-half the male cohort had average wage credits of less than $3,600 PYE, 1951-57. For only one group in the female cohort --age 55-59--was the proportion of workers with credits of $3,600 or more PYE, 1951-57, as high as 10 percent. Among women 25 to 69 years of age, the proportion with average credits of less than $1,200 PYE, 1951-57, varied from 35 percent to 45 percent. About 80 percent of all the women are to be found in the two lowest-paid cohorts--i.e., those with average credits of less than $2,400 PYE, 1951-57.

Analytical Concepts

Because the two time periods are of unequal duration, it was necessary to derive a common denominator by which employment and wage credits in each period could be compared. The Employment Ratio (ER) enables us to make such a comparison. It measures average intensity or continuity of employment of a cohort by relating the average number of years of work in employment covered by the Social Security Act to a given span of years. If a cohort, for example, was employed for an average of 7 years in covered employment during 1937-50, its ER was $.50\left[\frac{7 \text{ years employed}}{14 \text{ elapsed years}}\right]$. If the same cohort was also in covered employment an average of 7 years during 1951-57, its ER was $1.00\left[\frac{7 \text{ years employed}}{7 \text{ elapsed years}}\right]$. 5/ Thus, although the average number of years in covered employment of the hypothetical cohort was the same for both 1937-50 and 1951-57, the intensity of employment in the 1951-57 period was twice as great as in the 1937-50 period.

Average wage credits per year employed (PYE) indicate the average wage credits per year in covered employment during a given number of years. Average wage credits per elapsed year (PEY), on the other hand, are derived by dividing the average credits cumulated by the cohort by the number of elapsed years in a given time span. Average wage credits per elapsed year (PEY) thus reflect the joint effects of nonemployment 6/--as indicated by the ER--and by the level of wages PYE. As a welfare indicator, therefore, average wage credits PEY are a more significant measure than average credits PYE. Mathematically, the relationship between these three concepts may be expressed succinctly: Average wage credits PYE x ER = Average wage credits PEY.

1. Employment Intensity and Average Wage Credits, 1937-50, 1951-57

One of the principal uses of cohort analyses of wages is illustrated by the remainder of this paper. For wage earners in employment covered by the Social Security Act, we ask this question: To what degree was the increase in average earnings credits of the cohort due to increasing intensity or continuity of employment, and to what degree was this increase due to changes in the level of annual wage credits PYE? In the first part of this discussion, we examine employment changes from 1937-50 to 1951-57 as well as changes in the average wage credits per year employed and per elapsed year during these two periods. In the latter part, we examine the joint effects of changes in these two factors on average credits per elapsed year.

Male ER's (Table 3)

A comparison of the ER's for men for 1937-50 and 1951-57 indicates that as the male cohorts aged, intensity of employment (ER's) and the level of average annual wage credits PYE, 1951-57, on the whole were closely associated throughout the 1937-57 period. For those 50-54 years old in 1957, the ER's of the lowest-paid males were .407 in 1937-50; .671 in 1951-57; and .495 for 1937-57. For the intermediate low-paid males of this age, the ER's were correspondingly higher--.600 in 1937-50; .886 in 1951-57; and .695 for 1937-57. For the intermediate high-paid males in this age group the ER's for these three periods were still higher, .714, .943, and .790, respectively. For the highest-paid males of this age the ER's were highest--.864, .971, and .900, respectively. With the exception of the highest-paid males younger than 25 years of age, whose ER's probably reflect their more extensive schooling, ER's for 1937-50, 1951-57, and 1937-57 are closely associated with the level of average annual wage credits PYE, 1951-57.

Females--ER's (Table 4)

Among the female cohorts, the same general relationship between ER's and annual wage credits PYE, 1951-57, prevailed. With the exception of the small number of women in the highest-paid cohort--

5/ For workers aged 20 or younger in 1957 the ER was computed on the basis of years elapsing since 1937 or the year of birth, whichever was later.

6/ Nonemployment includes not only years of unemployment, but also periods which may involve withdrawal from the labor force or work in employment not covered under the Social Security Act.

MALE COHORTS

Table 3.--Employment Ratios (ER's) for 1937-50, 1951-57, and 1937-57, by Age in 1957 and Average Wage Credits Per Year Employed, 1951-57

Age in 1957	1937-50				1951-57				1937-57			
	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over
Total .	.129	.307	.493	.736	.529	.757	.886	.957	.262	.457	.624	.810
1-15	---	---	.643	---	.200	.329	.714	---	---	---	---	---
16-19	---	---	---	---	.329	.386	.357	.286	.131	.154	.143	.114
20-24	.014	.029	.043	.086	.614	.700	.771	.657	.214	.252	.286	.276
25-29	.157	.207	.271	.343	.614	.714	.814	.857	.310	.376	.452	.514
30-34	.300	.371	.407	.486	.643	.814	.871	.957	.414	.519	.562	.643
35-39	.364	.457	.514	.650	.686	.829	.886	.957	.471	.581	.638	.752
40-44	.414	.521	.643	.800	.671	.843	.914	.957	.500	.629	.733	.852
45-49	.400	.600	.693	.843	.671	.857	.929	.971	.490	.686	.771	.886
50-54	.407	.600	.714	.864	.671	.886	.943	.971	.495	.695	.790	.900
55-59	.450	.600	.750	.879	.657	.886	.943	.971	.519	.695	.814	.910
60-64	.429	.607	.757	.857	.671	.900	.957	.957	.510	.705	.824	.890
65-69	.386	.579	.779	.871	.714	.886	.957	.971	.495	.681	.838	.905
70 and over	.471	.614	.757	.814	.700	.871	.943	.929	.548	.700	.819	.852

FEMALE WAGE COHORTS

Table 4.--Employment Ratios (ER's) for 1937-50, 1951-57, and 1937-57, by Age in 1957 and Average Wage Credits Per Year Employed, 1951-57

Age in 1957	1937-50				1951-57				1937-57			
	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over
Total .	.186	.336	.507	.607	.543	.800	.914	.857	.305	.490	.643	.690
1-15	---	---	---	---	.186	.500	---	---	---	---	---	---
16-19	---	---	.050	---	.300	.357	.414	---	.120	.143	.206	---
20-24	.014	.014	.036	---	.600	.729	.786	---	.210	.252	.286	---
25-29	.164	.207	.250	.300	.629	.843	.929	.800	.319	.419	.476	.467
30-34	.307	.379	.464	.507	.600	.814	.929	.943	.405	.524	.619	.652
35-39	.336	.450	.586	.636	.571	.800	.914	.886	.414	.567	.695	.719
40-44	.314	.464	.600	.679	.600	.829	.914	.886	.410	.586	.705	.748
45-49	.300	.443	.600	.607	.614	.857	.914	.857	.405	.581	.705	.690
50-54	.271	.464	.607	.650	.629	.857	.914	.857	.390	.595	.710	.719
55-59	.279	.493	.614	.614	.629	.886	.929	.843	.395	.624	.719	.690
60-64	.257	.521	.650	.614	.643	.900	.929	.829	.386	.648	.743	.686
65-69	.257	.493	.629	.371	.686	.900	.857	.686	.400	.629	.705	.476
70 and over	.336	.486	.493	.779	.771	.929	.900	.843	.481	.633	.629	.800

less than 5 percent of the 21,000 women included in this study--ER's for 1937-50 and 1951-57 and average annual wage credits PYE, 1951-57, were also closely associated. For women 50-54 years old in 1957, ER's for the lowest-paid cohort were .271 in 1937-50; .629 in 1951-57; and .390 for 1937-57. For the intermediate low-paid women of this age, ER's were correspondingly higher--.464 in 1937-50; .857 in 1951-57, and .595 for 1937-57. For the intermediate high-paid women of this age, ER's were still higher--.607 for 1937-50; .914 for 1951-57; and .710 for 1937-57. For women with average credits of $3,600 or more PYE, 1951-57, ER's in 1937-50 were higher than for the three lower-paid female cohorts but were lower than those of the intermediate high-paid women in 1951-57. This variation is probably explained by differences in the employment propensity of women in the higher socio-economic levels.

Changes in ER's, 1937-50 to 1951-57 (Table 5)

The ER's furnish a dramatic summary of the patterns of increased employment which characterize the post-World War II economy, and of the differences between the male and female patterns of continuity of employment during this period.

While the male ER was higher than the female ER in both 1937-50 and 1951-57, Table 5 indicates that among the three lowest-paid cohorts, for both these periods the female ER exceeded the male ER. Furthermore, from 1937-50 to 1951-57 the ER's of women rose by 133 percent, while those of the male cohort increased by only 78 percent--a vivid reflection of the substantial rise in female employment during the post-World War II period.

For 1937-57 as a whole, among the three lowest-paid cohorts the ER's of the females were somewhat higher than those of the male cohorts; only for the highest-paid cohort was the female ER lower than the male ER.

Table 5.--ER's of Four Male and Female Cohorts as They Aged, 1937-50, 1951-57, and 1937-57

Sex and Wage Cohort	1937-1957	1937-1950	1951-1957	1951-1957 as % of 1937-1950
	(1)	(2)	(3)	(4)
Male cohorts	.567	.450	.800	177.8
Low-paid	.262	.129	.529	410.1
Intermediate low-paid.	.457	.307	.757	246.6
Intermediate high-paid	.624	.493	.886	179.7
High-paid	.810	.736	.957	130.0
Female cohorts	.433	.300	.700	233.3
Low-paid	.305	.186	.543	291.9
Intermediate low-paid.	.490	.336	.800	238.1
Intermediate high-paid	.643	.507	.914	180.3
High-paid	.690	.607	.857	141.2

Males--Average Wage Credits PYE (Table 6)

Average wage credits per year employed refer only to average credits of workers employed in jobs covered by the OASI program. Because they are calculated on a year-in-covered-employment basis, average credits PYE tend to be far more stable than average credits per elapsed year, which reflect trends in both the ER's and the average PYE's.

Although the level of average credits PYE generally increased from 1937-50 to 1951-57 for the male cohorts, the relative levels of average credits PYE of the four respective male cohorts remained unchanged during both time periods. The lowest-paid cohort during 1951-57 had the lowest average wage credits PYE for 1937-50. For each successively higher-paid cohort in 1951-57, average credits PYE for 1937-50 were also successively higher. Average wage credits PYE, 1951-57, of the four successively higher-paid male cohorts were $666, $1,819, $3,069, and $3,822, respectively. Average credits PYE, 1937-50, of the identical workers in these four cohorts were $684, $1,020, $1,422, and $2,048, respectively. These data suggest that the relative positions of the average low-paid and high-paid workers in the wage structure tend to be established at an early age and remain relatively unchanged throughout the working lifetime. It should be noted, however, that these are data for averages, and that the relative positions of substantial numbers of individual workers may improve with increased training or education during their working lifetimes.

As the members of the lowest-paid male cohort aged, the absolute level of the average wage credits per year employed of the older members--those aged 45 or older in 1957--declined from the 1937-50 levels. For those aged 55-59, for example, average credits PYE declined from $768 for 1937-50 to $694 for 1951-57. Similar declines occurred among workers in each of the older age groups in the lowest-paid cohort. Despite the rise in average annual wage credits PYE, which occurred among workers of every age in the three higher-paid cohorts, the data suggest that the annual wage credits per year employed of older, lower-paid workers tend to decline as they age, regardless of the secular upward trend in wage credits of all workers.

Females--Average Wage Credits PYE (Table 7)

The average wage credits per year employed of the female cohorts differ from those of the male cohorts in one significant respect. Although the upward trend which characterized the male average PYE from 1937-50 to 1951-57 also characterized the female average PYE, the older lowest-paid female cohort--unlike the older lowest-paid male cohort--suffered no decline in average wage credits PYE. From 1937-50 to 1951-57 average credits per year employed of each age group (except for those 70 or older) in the lowest-paid female cohort increased. Among the lowest-paid women aged 40-44, for example, average credits PYE rose from $593 for 1937-50 to $708 for 1951-57, or 19 percent. For women aged 50-54, average credits PYE increased

MALE WAGE COHORTS

Table 6.--Average Wage Credits Per Year Employed (PYE) for 1937-50, 1951-57, and 1937-57, by Age in 1957 and Average Wage Credits Per Year Employed, 1951-57

Age in 1957	1937-50				1951-57				1937-57			
	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over
Total ..	$684	$1,020	$1,442	$2,048	$666	$1,819	$3,069	$3,822	$672	$1,461	$2,212	$2,747
1-15	---	---	1,782	---	245	1,572	3,481	---	245	1,572	2,389	---
16-19	---	---	---	---	510	1,575	2,729	4,003	510	1,575	2,729	4,003
20-24	96	180	465	476	704	1,744	2,835	3,714	677	1,626	2,598	3,044
25-29	425	618	833	1,104	753	1,828	3,018	3,784	642	1,384	2,144	2,593
30-34	549	806	1,057	1,432	742	1,856	3,111	3,806	649	1,355	2,119	2,610
35-39	616	861	1,204	1,669	701	1,885	3,143	3,820	657	1,348	2,101	2,581
40-44	646	985	1,402	1,941	688	1,848	3,103	3,828	665	1,371	2,109	2,647
45-49	711	1,086	1,560	2,194	685	1,875	3,083	3,830	699	1,415	2,171	2,792
50-54	715	1,174	1,702	2,316	695	1,857	3,080	3,830	706	1,464	2,250	2,861
55-59	768	1,226	1,882	2,435	694	1,869	3,099	3,831	737	1,499	2,352	2,932
60-64	935	1,350	1,923	2,465	716	1,857	3,123	3,824	839	1,566	2,388	2,952
65-69	890	1,456	2,049	2,508	732	1,825	3,047	3,812	814	1,616	2,429	2,975
70 and over	1,274	1,685	2,076	2,541	694	1,767	3,030	3,813	1,027	1,719	2,442	3,003

FEMALE WAGE COHORTS

Table 7.--Average Wage Credits Per Year Employed (PYE) for 1937-50, 1951-57, and 1937-57, by Age in 1957 and Average Wage Credits Per Year Employed, 1951-57

Age in 1957	1937-50				1951-57				1937-57			
	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over
Total .	$574	$901	$1,282	$1,852	$653	$1,764	$2,929	$3,800	$621	$1,370	$2,063	$2,658
1-15	---	---	---	---	174	1,877	---	---	174	1,877	---	---
16-19	---	---	1,126	---	467	1,542	2,623	---	467	1,540	2,332	---
20-24	196	118	366	---	658	1,708	2,710	---	637	1,648	2,522	---
25-29	464	696	889	1,128	662	1,776	2,875	3,741	594	1,420	2,180	2,621
30-34	590	879	1,264	1,587	667	1,809	2,956	3,750	628	1,361	2,110	2,629
35-39	561	790	1,188	1,595	657	1,774	2,960	3,780	605	1,253	1,965	2,492
40-44	593	822	1,209	1,630	708	1,753	2,960	3,790	649	1,261	1,966	2,483
45-49	597	928	1,326	1,867	694	1,794	2,947	3,812	646	1,354	2,027	2,672
50-54	578	940	1,368	1,944	699	1,773	2,944	3,812	643	1,340	2,045	2,686
55-59	548	1,002	1,463	2,125	705	1,792	2,940	3,821	631	1,376	2,099	2,815
60-64	636	1,097	1,534	2,227	674	1,801	2,933	3,823	657	1,423	2,117	2,870
65-69	656	1,147	1,593	2,139	705	1,721	2,886	3,881	684	1,421	2,117	2,975
70 and over ...	718	1,279	1,767	2,462	690	1,666	2,934	3,866	703	1,468	2,324	2,955

from $578 to $699, a rise of 21 percent. Even among women aged 65-69, average credits PYE rose, increasing from $656 for 1937-50 to $705 for 1951-57, or 8 percent.

Despite the substantial percentage increase in the average credits per year employed of the lowest-paid women, the low levels which characterized the average PYE of these women in both periods indicate that when employed, they were in the main part-year workers. The low levels of the average PYE for both periods also suggest that to a considerable degree the increased employment of women in the postwar period involves a rise in the number of part-year workers. While the average credits PYE of the two higher-paid female cohorts who, it might be inferred, number a larger proportion of full-year workers, also rose, the proportion of high-paid women was relatively small. Since four-fifths of the women were in the two lowest-paid cohorts, the data suggest that for most working women, employment and wages are still second in importance to their responsibilities as wives and mothers.

Changes in Average Wage Credits PYE, 1937-50 to 1951-57 (Table 8)

The relatively greater percentage increase in the ER's of women as compared to those of men from 1937-50 to 1951-57 was accompanied by a greater relative increase in the female average wage credits PYE. The average annual wage credits PYE of the males increased by 67 percent from 1937-50 to 1951-57; for the females, the average PYE increased by 73 percent. Indeed, for the lowest-paid cohort, the male average PYE declined about 3 percent, while average credits PYE for the lowest-paid women rose by 14 percent. For the three highest-paid cohorts, the increase in average wage credits PYE was greater for women than for men, rising by 96 percent for intermediate low-paid women as compared to 78 percent for men in the same wage cohort; by 129 percent for women as compared to 113 percent for men in the intermediate high-paid cohort; and by 105 percent for women, but only by 87 percent for men in the highest-paid cohort.

Table 8.--Average Wage Credits Per Year Employed of Four Male and Female Cohorts as They Aged, 1937-50, 1951-57, and 1937-57

Sex and Wage Cohort	1937-1957	1937-1950	1951-1957	1951-1957 as % of 1937-1950
	(1)	(2)	(3)	(4)
Male cohorts	$2,181	$1,658	$2,770	167.1
Low-paid	672	684	666	97.4
Intermediate low-paid.	1,461	1,020	1,819	178.3
Intermediate high-paid	2,212	1,442	3,069	212.8
High-paid	2,747	2,048	3,822	186.6
Female cohorts	1,373	989	1,706	172.5
Low-paid	621	574	653	113.8
Intermediate low-paid.	1,370	901	1,764	195.8
Intermediate high-paid	2,063	1,282	2,929	228.5
High-paid	2,658	1,852	3,800	205.2

Males--Average Wage Credits PEY, 1937-50 and 1951-57 (Table 9)

The discussion thus far has demonstrated that as workers aged from 1937 to 1957 there was a general increase in intensity of employment, as measured by the ER, and in the level of average annual wage credits per year employed. The joint effects of these two factors may be discerned in the average wage credits per elapsed year.

For the four male cohorts, Table 9 indicates that when the cohorts are defined on the basis of the level of average wage credits PYE for 1951-57 there is a close association between the relative levels of average wage credits PYE and average wage credits PEY for both the 1937-50 and 1951-57 periods, and consequently for the entire 21-year time span, 1937-57. For each of the four male cohorts average wage credits PEY for 1951-57 varied directly with their respective average wage credits PYE for 1951-57. The levels of the average credits PEY for the four male cohorts for 1951-57 were also consistent with the respective levels of the average credits PEY for these same cohorts for 1937-50--i.e., the highest-paid cohort in 1951-57 had the highest average wage credits PEY for both 1951-57 and 1937-50; the lowest-paid cohort in 1951-57 had the lowest average wage credits PEY, 1937-50 and 1951-57. The relative average wage credits PEY for 1937-50 and 1951-57 for the two intermediate cohorts were in an intermediate position in both time periods.

An interesting instance of the difference between the average PYE and the average PEY, and of its significance, may be found when we compare the two types of wage credits for those aged 55-59 in the lowest-paid male cohort for 1937-50 and 1951-57, as follows:

Type of wage credit	1937-50	1951-57
Average PYE	$768	$694
Average PEY	$346	$460

What is true for the 55-59 year-old group in the lowest-paid male cohort is true for each of the groups aged 40 to 69 in this cohort. Although the level of their average wage credits PYE declined as they aged from 1937-50 to 1951-57, the intensity of their employment--i.e., their ER--increased to such an extent that their average credits per elapsed year increased. Thus, the increase in employment resulted in an increase in average wage credits PEY, despite the decline in their average wage credits PYE.

MALE WAGE COHORTS

Table 9.--Average Wage Credits Per Elapsed Year (PEY) for 1937-50, 1951-57, and 1937-57, by Age in 1957 and Average Wage Credits Per Year Employed, 1951-57

Age in 1957	1937-50				1951-57				1937-57			
	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over
Total ..	$88	$313	$711	$1,507	$352	$1,390	$2,701	$3,666	$176	$672	$1,374	$2,227
1-15	---	---	1,146	---	49	524	2,486	---	---	---	---	---
16-19	---	---	---	---	170	606	964	1,144	---	---	---	---
20-24	1	5	20	41	431	1,213	2,189	2,449	145	408	743	843
25-29	67	128	226	379	463	1,317	2,469	3,233	199	524	974	1,330
30-34	165	299	430	696	479	1,516	2,715	3,638	269	705	1,192	1,677
35-39	224	393	619	1,085	483	1,565	2,774	3,649	310	784	1,337	1,940
40-44	267	513	901	1,553	463	1,558	2,834	3,680	333	861	1,546	2,262
45-49	284	652	1,081	1,850	463	1,617	2,872	3,739	344	973	1,678	2,479
50-54	291	704	1,215	2,001	470	1,640	2,904	3,719	351	1,016	1,778	2,574
55-59	346	736	1,411	2,140	460	1,656	2,929	3,742	384	1,042	1,917	2,674
60-64	401	819	1,456	2,112	481	1,673	2,990	3,681	428	1,104	1,967	2,635
65-69	344	843	1,596	2,184	528	1,617	2,913	3,681	405	1,101	2,035	2,683
70 and over	600	1,035	1,571	2,068	408	1,535	2,854	3,554	563	1,202	1,999	2,564

FEMALE WAGE COHORTS

Table 10.--Average Wage Credits Per Elapsed Year (PEY) for 1937-50, 1951-57, and 1937-57, by Age in 1957 and Average Wage Credits Per Year Employed, 1951-57

Age in 1957	1937-50				1951-57				1937-57			
	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over	Under $1,200	$1,200-$2,399	$2,400-$3,599	$3,600 and over
Total ..	$107	$303	$650	$1,124	$357	$1,421	$2,666	$3,268	$190	$675	$1,322	$1,839
1-15	---	---	---	---	32	939	---	---	---	---	---	---
16-19	---	---	56	---	141	543	1,087	---	---	---	---	---
20-24	3	2	13	---	395	1,236	2,146	---	133	413	724	---
25-29	76	144	222	338	413	1,508	2,669	2,993	189	599	1,038	1,223
30-34	181	333	586	805	400	1,466	2,764	3,514	254	711	1,312	1,708
35-39	188	355	696	1,014	376	1,426	2,720	3,363	251	712	1,371	1,797
40-44	186	381	725	1,107	429	1,463	2,706	3,379	267	742	1,386	1,864
45-49	179	411	796	1,133	431	1,541	2,681	3,267	263	788	1,424	1,845
50-54	157	436	830	1,264	438	1,515	2,672	3,243	250	796	1,444	1,923
55-59	153	494	898	1,305	443	1,578	2,720	3,209	250	855	1,505	1,940
60-64	163	572	997	1,367	436	1,637	2,720	3,139	254	927	1,571	1,958
65-69	169	565	1,002	794	489	1,553	2,499	2,619	275	895	1,501	1,402
0 and over	241	622	871	1,918	529	1,537	2,650	3,223	337	927	1,464	2,353

Females--Average Wage Credits PEY, 1937-50 and 1951-57 (Table 10)

Average wage credits per elapsed year of the four female cohorts increased for each age group from 1937-50 to 1951-57. While the average PEY of the lowest-paid female cohort increased about three and a half times from 1937-50 to 1951-57, the level of the average PEY for these women, even in 1951-57 is still quite low. This reinforces our conclusion that the increased employment of women in the post-World War II period in the main involves an increase in the number of women who are employed on a part-year basis and whose employment patterns tend to be intermittent, or perhaps in many instances, casual in nature.

Changes in Average Wage Credits PEY, 1937-50 to 1951-57 (Table 11)

The joint effects of increasing levels of employment and annual wage credits per year employed are reflected in the changes in average wage credits PEY from 1937-50 to 1951-57. Average credits per elapsed year of the females increased at a faster rate than those of the males, rising by 302 percent for females, but by only 197 percent in the case of the males. Although the average PEY increased among all the four cohorts, it varied substantially by average PYE level, increasing most among the two lowest-paid male cohorts, and among the two intermediate-paid female cohorts. For the males, the largest increase in average credits per elapsed year occurred among the intermediate low-paid cohort, rising by 339 percent for this cohort, and by 300 percent for the lowest-paid cohort. Average credits per elapsed year of the highest-paid cohorts increased by about 282 and 142 percent, respectively.

Table 11.--Average Wage Credits Per Elapsed Year of Four Male and Female Cohorts as They Aged, 1937-50, 1951-57, and 1937-57

Sex and Wage Cohort	1937-1957	1937-1950	1951-1957	1951-1957 as % of 1937-1950
	(1)	(2)	(3)	(4)
Male cohorts	$1,237	$746	$2,216	297.1
Low-paid	176	88	352	400.0
Intermediate low-paid.	668	313	1,377	439.9
Intermediate high-paid	1,380	711	2,719	382.4
High-paid	2,225	1,507	3,658	242.7
Female cohorts	595	297	1,194	402.0
Low-paid	189	107	355	331.8
Intermediate low-paid.	671	303	1,411	465.7
Intermediate high-paid	1,327	650	2,677	411.8
High-paid	1,834	1,124	3,257	289.8

For the women, average credits per elapsed year increased most among the two intermediate cohorts, rising by 366 percent and 312 percent in the case of the intermediate low-paid and intermediate high-paid cohorts, respectively. Average credits per elapsed year of the lowest-paid women increased by 232 percent, and by 190 percent in the case of the highest-paid women.

2. Joint Effects of Changes in ER and PYE on Changes in PEY, 1937-50 to 1951-57

Use of cohort analysis permits us to evaluate the relative effects of changes in intensity of employment and in the level of average credits PYE on changes in average credits per elapsed year. Put somewhat differently, this means that the joint effects of changes in employment and wage credit levels on the movement of average wage credits per elapsed year can be disentangled, and the relative importance of each of these changes can be measured.

The data in Tables 12 and 13 indicate that the impact of changes from 1937-50 to 1951-57 in these two factors on changes in average credits per elapsed year differed for each sex, but for both the male and female cohorts, these changes were age-related. For workers in each of the four wage cohorts, however, the relative impact of changes in ER's and average wage credits PYE on changes in average credits PEY was approximately the same.

Males--Effect of Changes in ER's and Average Wage Credits PYE (Table 12)

From 1937-50 to 1951-57, the ER's of each of the four male cohorts increased substantially, but the extent of the increase varied inversely with the level of average wage credits PYE, 1951-57. Average credits PYE, however, did not rise in the same fashion. For the three highest-paid male cohorts, the percent increase in the average PYE, by age group, was greatest for the intermediate high-paid cohort, of somewhat lesser proportion for the highest-paid cohort, and was smallest for the intermediate low-paid cohort. For the lowest-paid cohort, average wage credits PYE did not increase, but declined from 1937-50 to 1951-57 for cohort members aged 45 or older.

The joint effect of these changes is reflected in the changes in average credits per elapsed year. For each of the male cohorts, average credits PEY increased substantially from 1937-50 to 1951-57. For the three lowest-paid cohorts the extent of the increase varied directly with the level of average credits PYE; for the highest-paid cohort, the percentage increase in average credits PEY was greater than the increase in the average credits PEY of the lowest-paid cohort, but of lesser proportions than the increase in the average PEY's of the two intermediate-paid cohorts.

In general, the percentage increase in ER's and average credits PYE varied inversely with age among the three higher-paid cohorts--the older the age group, the lower was the percent increase

MALE COHORTS

Table 12.--ER's, Average Wage Credits PYE and PEY--1951-57 Levels as a Percent of 1937-50 Levels

Age in 1957	Average Wage Credits PYE, 1951-57															
	Under $1,200				$1,200-$2,399				$2,400-$3,599				$3,600 and Over			
	ER	PYE	PEY	Ratio of (1) to (2)	ER	PYE	PEY	Ratio of (1) to (2)	ER	PYE	PEY	Ratio of (1) to (2)	ER	PYE	PEY	Ratio of (1) to (2)
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Total.	410.1	97.4	400.0	4.21	246.6	178.3	444.1	1.38	179.7	212.8	379.9	.84	130.0	186.6	243.3	.70
1-15 ..	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
16-19 ..	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
20-24 ..	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
25-29 ..	391.1	177.2	---	2.21	344.9	295.8	---	1.17	300.4	362.3	---	.83	249.9	342.8	---	.73
30-34 ..	214.3	135.2	290.3	1.59	219.4	230.3	507.0	.95	214.0	294.3	631.4	.73	196.9	265.8	522.7	.74
35-39 ..	188.5	113.8	215.6	1.66	181.4	218.9	398.2	.83	172.4	261.0	448.1	.66	147.2	228.9	336.3	.64
40-44 ..	162.1	106.5	173.4	1.52	161.8	187.6	303.7	.86	142.1	221.3	314.5	.64	119.6	197.2	237.0	.61
45-49 ..	167.8	96.3	163.0	1.74	142.8	172.7	248.0	.83	134.1	197.6	265.7	.68	115.2	174.6	202.1	.66
50-54 ..	164.9	97.2	161.5	1.70	147.7	158.2	233.0	.93	132.1	181.0	239.0	.73	112.4	165.4	185.9	.68
55-59 ..	146.0	90.4	132.9	1.62	147.7	152.4	225.0	.97	125.7	164.7	207.6	.76	110.5	157.3	174.9	.70
60-64 ..	156.4	76.6	120.0	2.04	148.3	137.6	204.3	1.08	126.4	162.4	205.4	.78	111.7	155.1	174.3	.72
65-69 ..	185.0	82.2	153.5	2.25	153.0	125.3	191.8	1.22	122.8	148.7	182.5	.83	111.5	152.0	168.5	.73
70 and over ..	148.6	54.5	81.3	2.73	141.9	104.9	148.3	1.35	124.6	146.0	181.7	.85	114.1	150.1	171.9	.76

FEMALE COHORTS

Table 13.--ER's, Average Wage Credits PYE and PEY--1951-57 Levels as a Percent of 1937-50 Levels

Age in 1957	Average Wage Credits PYE, 1951-57															
	Under $1,200				$1,200-$2,399				$2,400-$3,599				$3,600 and Over			
	ER	PYE	PEY	Ratio of (1) to (2)	ER	PYE	PEY	Ratio of (1) to (2)	ER	PYE	PEY	Ratio of (1) to (2)	ER	PYE	PEY	Ratio of (1) to (2)
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Total.	291.9	113.8	333.6	2.57	238.1	195.8	469.0	1.22	180.3	228.5	410.2	.79	141.2	205.2	290.7	.69
1-15 ..	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
16-19 ..	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
20-24 ..	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
25-29 ..	383.5	142.7	---	2.69	407.2	255.2	---	1.60	371.6	323.4	---	1.15	226.7	331.6	---	.80
30-34 ..	195.4	113.1	221.0	1.73	214.8	205.8	440.2	1.04	200.2	233.9	471.7	.86	186.0	236.3	436.5	.79
35-39 ..	169.9	117.1	200.0	1.45	177.8	224.6	401.7	.79	156.0	249.2	390.8	.63	139.3	237.0	331.7	.59
40-44 ..	191.1	119.4	230.6	1.60	178.7	213.3	384.0	.84	152.3	244.8	373.2	.62	130.5	232.5	305.2	.56
45-49 ..	204.7	116.2	240.8	1.76	193.5	193.3	374.9	1.00	152.3	222.2	336.8	.69	141.2	204.2	288.3	.69
50-54 ..	232.1	120.9	279.0	1.92	184.7	188.6	347.5	.98	150.6	215.2	321.9	.70	131.8	196.1	256.6	.67
55-59 ..	225.4	128.6	289.5	1.75	179.7	178.8	319.4	1.01	151.3	201.0	302.9	.75	137.3	179.8	245.9	.76
60-64 ..	250.2	106.0	267.5	2.36	172.7	164.2	286.2	1.05	142.9	191.2	272.8	.75	135.0	171.7	229.6	.79
65-69 ..	266.9	107.5	289.3	2.48	182.6	150.0	274.9	1.22	136.2	181.2	249.4	.75	184.9	181.4	329.8	1.02
70 and over ..	229.5	96.1	219.5	2.39	191.2	130.3	247.1	1.47	182.6	166.0	304.2	1.10	108.2	157.0	168.0	.69

in their ER's and average wage credits PYE. For the lowest-paid cohort, the pattern of change was also inversely related to age, but in this instance, the level of average credits PYE of cohort members aged 45 or older declined from 1937-50 to 1951-57.

These findings are stated abstractly, because of the limitations of time. A brief description of changes in the employment and wage levels of workers aged 55-59, and of the joint effects of the changes in these factors on average wage credits PEY, may serve to illustrate in a more concrete fashion some of the uses to which this mode of analysis may be used.

For men aged 55-59 in the lowest-paid cohort, average credits per elapsed year increased by 33 percent from 1937-50 to 1951-57. Their average credits per year employed, however, declined by 10 percent during this period, but the decline was more than offset by an increase of 46 percent in their ER's. For these workers, therefore, the increase in ER's had a measurably greater impact on changes in average credits per elapsed year than did the changes in the level of average credits PYE. The ratio of changes in ER's to changes in the average PYE indicates that the ER factor was 62 percent more significant than the PYE factor in causing the increase in average credits per elapsed year. The increase in employment intensity from 1937-50 to 1951-57, therefore, was the major factor responsible for the rise in the average credits PEY of the lowest-paid men in this age group.

For men aged 55-59 in the highest-paid cohort, the increase in average credits per elapsed year must be explained by an entirely different relationship between the ER and PYE components. Although average credits per elapsed year of this group increased by 75 percent from 1937-50 to 1951-57, the primary cause of the increase is to be found in the change in average credits PYE. ER's of this group increased by only 11 percent, but average credits PYE rose by 57 percent. While the joint effect of these two changes resulted in a rise of 75 percent in average wage credits PEY, the "change" ratio indicates that for this age group, ER changes were only 70 percent as significant as the changes in the level of average credits in causing the increase in average credits per elapsed year. Thus, in the case of the lowest-paid males aged 55-59, the increase in employment intensity was the sole cause of the increase in the average wage credits PEY; for the highest-paid cohort, the increase in employment intensity was responsible only to a minor degree for the significant increase in average credits per elapsed year.

Females--Effect of Changes in ER's and PYE's (Table 13)

For the women cohorts, the relationship between changes in ER's and average credits PYE is more complex. Because of the intermittency of employment, the female patterns of change are less regular than the male patterns, and the ever-present time limitation requires that we confine our discussion to a single age group--women aged 55 to 59 years of age. The data for all the age groups in each of the cohorts may be examined in detail in Table 13.

For women aged 55-59, the ER of the lowest-paid cohort rose by 125 percent from 1937-50 to 1951-57. During this period, the level of average credits PYE of this age group increased by only 29 percent. The joint effect of these patterns resulted in an increase of 190 percent in their average credits per elapsed year. Although the average PYE rose substantially, the ratio of ER to average PYE changes indicates that the increase in employment intensity was one and three-fourths times more significant than the increase in the level of average credits PYE in causing the rise in average credits per elapsed year.

For the highest-paid women in this age group--as among the highest-paid men--the relative significance of the two components of PEY was reversed. ER's of the highest-paid female cohort increased by 37 percent from 1937-50 to 1951-57, but average wage credits PYE rose by 80 percent during this time. Thus, although average credits of the highest-paid women rose by 146 percent, the increase in ER's was only 76 percent as important as the increase in the level of average credits PYE in accounting for the rise in average credits per elapsed year.

SAMPLE SIZE DETERMINATION IN SIMPLE ANALYTICAL SURVEYS (Abstract)

Joseph Sedransk, Harvard University

The primary purpose of an "analytic" sample survey is to compare various sectors (strata, groups) of some population. Considering the case where the units are classified by only one variable, the comparisons can be made in many ways. If one can sample independently within each of the groups, a survey may be designed which ensures that the selected form of comparison will be made in some "best" manner.

If it is impossible to sample independently in each of the groups which are to be compared, an analytic survey may still be designed to make the same sort of comparisons as in the straight-forward case. A double sampling procedure which provides that the comparisons will be made with a specified precision "on the average" may be devised. A sequential sampling scheme is also available. It is required that the variance of the difference for each pair of group means be no larger than a specified constant. The probability that the precision specifications are satisfied by stage n in the sampling is determined.

THE LIFE TABLE FROM THE STANDPOINT OF STATISTICAL ANALYSIS (Abstract)

Norman B. Ryder, University of Wisconsin

In this paper, attention is given to the form of the life table from the standpoint of its usefulness in an analytic rather than actuarial context. The points made in the paper revolve around a new measure, the mean length of time spent within an age interval by those who die within that interval, symbolized by ${}_na_x$. Use of this measure yields a solution to the paedagogical problems posed by the conventional life table. The life table columns are examined from the standpoint of their relationships to the standard statistical procedures for representing a frequency distribution of ages at death. By employment of ${}_na_x$, the network of formal relationships among various mortality and stationary population elements is able to be presented in elementary algebraic form. On the basis of an examination of the research purposes of the life table, a case is made for using ${}_nm_x$ rather than ${}_nq_x$, and for deleting ${}_nd_x$, ${}_np_x$, and e_x. The most convenient mode of manufacture of the life table is based on data for ${}_nm_x$ and data or assumptions about ${}_na_x$. Sources of error in the former are discussed from the standpoint of the basic data. Modes of summarization of mortality are discussed from the standpoint of separating the amount and the time components of mortality, and distinguishing the positive and the negative forms of statement. The paper is concluded with some suggestions for simple construction of a life table.

A COMPARISON ON THE BASIS OF THE NASHVILLE MORBIDITY SURVEY OF TWO FRAMES WITH EQUAL AND UNEQUAL SIZED FIRST-STAGE UNITS

*T. Khosla, University of Aberdeen
A. L. Finkner, Research Triangle Institute,
and
J. C. Koop, N. C. State College

1. Introduction

In any statistical investigation where it may be costly to employ unrestricted random sampling and where units that are close together exhibit little variability, multi-stage sampling is appropriate. The advantages of multi-stage sampling are reduced cost both in travelling, because the ultimate units are less dispersed, and in developing the frame since it is constructed in successive stages.

However, with multi-stage sampling the question of whether to select the first-stage units with equal or unequal probability arises. Now the selection of the first-stage units with equal probability is usually less efficient than the corresponding selection with unequal probability if the first-stage units are large and vary greatly in their sizes. Difficulties of unequal probability sampling without replacement are found in (1) involved variance formulas with complicated functions of the probabilities used in the selection of the units and (2) estimates of variance which may be negative.

These reasons have prompted us to make a comparative study of a frame in which the first-stage units are constructed of equal size (in terms of ultimate sampling units) and selected with equal probabilities and without replacement, with another frame in which the first-stage units are unequal and selected with probability proportional to the size of the ultimate sampling units and also without replacement. Clearly when first stage units are equal in size equal probability sampling and sampling with probability proportional to size are equivalent.

The two methods will be compared on the basis of relative precision and also with an eye on relative costs. Precision will be judged from the closeness with which estimates centre round their own mean, in repeated application of the same sampling scheme.

2. The Nashville Morbidity Survey

The data used in the present study were obtained from the Nashville Morbidity Survey (N.M.S.), a detailed report of which is given by Finkner et al. (1960). In this section a brief description of this survey will be given with particular reference to (1) the construction of its frame and (2) the method of sampling. These aspects of the original study have an important bearing on the reconstruction of the frame for the purpose of the present study.

The universe consisted of eligible households found in the 51 city planning units of the city of Nashville, Tennessee, and its environs. Each of these planning units were divided up into a certain number of strata. Each stratum was further subdivided into 45 sampling units. The sampling unit was defined as an area segment, or as a strip along a street, with an expectation of about one household per sampling unit. The total number of dwelling units included in the universe at the time the frame was constructed was 69,244.

The structure of this frame which is based on maps is shown in the following table.

Table 1. The Structure of the Frame of the Nashville Morbidity Survey

Planning unit no.	Number of dwelling units	Number of strata	Number of sampling units	"Expected" size of sampling unit
I	II	III	IV	V
12	886	19	855	1.036
13	999	22	990	1.009
14	2287	50	2250	1.016
15	1331	29	1305	1.020
16	1731	38	1710	1.012
17	2268	50	2250	1.008
18	1605	35	1575	1.019
19	1171	26	1170	1.001
20	1003	22	990	1.013
21	242	5	225	1.076
22	1800	40	1800	1.000
23	1208	26	1170	1.032
24	1636	36	1620	1.010
25	1223	27	1215	1.007
26	1130	25	1125	1.004
27,28,29	939	20	900	1.043
30	2945	65	2925	1.007
31	2203	48	2160	1.020
32	1863	41	1845	1.010
33	1771	39	1775	1.009
35	1861	41	1845	1.009
34,36	1940	43	1935	1.003
37	1730	38	1710	1.012
38	684	15	675	1.013
39	152	3	135	1.126
40	487	10	450	1.082
41	340	7	315	1.079
42	2313	51	2295	1.008
43	2921	64	2880	1.014

* Formerly at North Carolina State College

Table 1. (continued)

44	938	20	900	1.042
45,46	1423	31	1395	1.020
47	939	20	900	1.043
48	2791	62	2790	1.000
49	2529	56	2520	1.004
50	3782	84	3780	1.001
51	1931	42	1890	1.022
52	463	10	450	1.029
53	210	4	180	1.167
54	1080	24	1080	1.000
55	934	20	900	1.038
56	1049	23	1035	1.014
57	575	12	540	1.065
65	1467	32	1440	1.019
66	1614	35	1575	1.025
67	1768	39	1755	1.007
73	1914	42	1890	1.013
74	1148	25	1125	1.020
Total	69,224	1516	68,220	1.015

In column I of Table 1 the serial number of the planning unit is given. Not all of the planning units were included in the survey. Thus planning unit number 12, containing 886 dwelling units (existing about the time the frame was constructed), was divided up into 19 strata each of which was further subdivided into 45 sampling units (each an area segment) giving in all 855 sampling units shown in column IV. It will be noted that all the entries in this column are multiples of 45. The "expected" size of the sampling unit, shown in column V, is 886/855, i.e. a little over one household. In all there were 1516 strata.

In regard to the method of sampling, two sampling units (area segments) were selected with equal probability and without replacement from the forty-five units constituting each of the 1516 strata. This observation has an important bearing on the statements on the structure of the sample design of the present study made in Section 4. Also it will be noted that the sample was self-weighting.

Out of 2 x 1516 = 3032 sampling units selected 282 were found to have no dwelling units. In the remaining sampling units, 2649 completed interviews from households were obtained. This number included 85 out of 264 households selected at random which did not respond at the first, but which responded at the second or the third interview*. The details of all these complexities in the resulting data can be found in the report referred to above. At any rate, the present study was not noticeably affected by these complexities in the data.

*Vaivanijkul (1961) studied the differences between respondents and these initial non-respondents for 34 morbidity characteristics of the survey. She found significant differences in only six characteristics.

3. Reconstruction of Frames for Present Study

In this section, the procedure of forming two new frames from existing materials shown in Table 1 will be given. Consider the universe as defined by the 51 planning units of the N.M.S. as shown in column I of Table 1. We desire only 5 large strata each composed of an equal number of large first-stage units. There are 51 planning units. Deleting no. 17 and no. 23 (for reasons which will appear in the subsequent discussion) and compounding the contiguous units 27, 28 and 29 (all of which contain only 939 dwelling units) as one unit, also 34 and 36 as one, and finally 45 and 46 as one, leaves us with the original 42 planning units and the 3 new compounded units whose respective constituents are contiguous. Thus in all we have 45 units and these we redefine as first-stage units (FSU's) and the original strata of the N.M.S. which make up each of these new units as second-stage units (SSU's), and lastly the sampling units (area segments) which the original N.M.S.-strata contained as the third-stage units.

Frame A. Dividing up the 45 FSU's into contiguous sets of 9, each to constitute a stratum, we have Frame A as shown in Table 2.

It will be noted that the nine FSU's in each stratum are quite unequal in size as measured by the number of SSU's contained by each of them. The FSU's are selected with probability proportional to size, e.g. the selection probability for FSU no. 1 stratum I is
19/(19+22+50+29+39+35+26+22+5).

Table 2. Frame A[a/]

Strata	Planning unit no. of N.M.S.	FSU no. of Frame A	Number of SSU's in Frame A
	12	1	19
	13	2	22
	14	3	50
	15	4	29
I	16	5	38
	18	6	35
	19	7	26
	20	8	22
	21	9	5
	22	1	40
	24	2	36
	25	3	27
	26	4	25
II	27,28,29	5	20
	30	6	65
	31	7	48
	32	8	41
	33	9	32
	35	1	41
	34,36	2	43
	37	3	38
	38	4	15
III	39	5	3
	40	6	10
	41	7	7

	42	8	51
	43	9	64
	44	1	20
	45,46	2	31
	47	3	20
	48	4	62
IV	49	5	56
	50	6	84
	51	7	42
	52	8	10
	53	9	4
	54	1	24
	55	2	20
	56	3	23
	57	4	12
V	65	5	32
	66	6	35
	67	7	39
	73	8	42
	74	9	25

[a] Each SSU contains 45 sampling units

We shall now construct another frame (on the basis of the same material) consisting of 5 strata in which the FSU's are of equal size as measured by the number of SSU's.

Frame B. In this frame, our object is to construct FSU's of equal size. With the deletion of two planning units (no. 17 and no. 23) from the original N.M.S. frame, we are left with 1440 SSU's. These are equally allotted to the 45 FSU's, so that each FSU contains exactly 32 SSU's under this scheme.

Consider Frame A. The first FSU (no. 12) has 19 SSU's and with 13 consecutive SSU's taken from the second FSU (no. 13) and added to it the size of the first FSU comes to 32. This leaves the second FSU with 9 SSU's, which in turn is compensated by taking 23 consecutive SSU's from the third FSU (no. 14) and the process is repeated in like manner for the remaining units. This procedure is diagrammatically represented for Stratum I in Table 3. After redefining the new first-stage units as above, 5 strata are formed, each stratum consisting of 9 first-stage units taken in order.

Here again, it may be mentioned that a number of different combinations can be made to form the first-stage units of equal size. For example, 13 second-stage units are required to be added to the size of the first FSU. Theoretically, this can be done in a number of ways. We may take 13 SSU's in any arbitrary manner from the remaining FSU's. Geographical proximity and administrative conveniences were again the guiding factors.

All these FSU's are selected with equal probabilities.

Table 3. Frame B (Stratum I)[a]

Planning unit no. of N.M.S.	Number of strata in N.M.S.	Number of SSU's in Frame B		FSU no. of Frame B
12	19	19	32	1
		13		
13	22	9	32	2
		23		
14	50	27	32	3
		5		
15	29	24	32	4
		8		
16	38	30	32	5
		2		
18	35	32	32	6
		1		
19	26	26	32	7
		5		
		17		
20	22	5	32	8
		10		
21	5			
22	40	30	32	9
		2		
24	36	32		

[a] Here also each SSU contains 45 sampling units

4. The Samples for Present Study

In the present study, two first-stage units are selected without replacement from each stratum, but with probability proportional to size of first-stage unit in sampling from Frame A, and with equal probability in sampling from Frame B. Now as the FSU's are composed of the 1440 SSU's, which were formerly strata of the N.M.S., all information is already available and are recorded on cards. Thus in any given stratum of Frame A or B, all the SSU's constituting each selected FSU play the role of second-stage units which are completely "sampled" and the two sampling units selected out of 45 from the original strata now play the role of the third-stage units under the present system of dual reconstruction of the original frame for the purpose of comparison.

The situation now is as if two different sample surveys relating to the same universe had been carried out with two different types of frames described in Section 3.

5. Theoretical Basis for Comparisons

We now introduce the formulas for estimates and their variances and also discuss methods for comparing the precision of the estimates obtained

on the basis of Frames A and B. The formulas will be given for a single stratum to avoid making the notation more involved than it is now.

Let Y_{iks} be the measure of a character of interest in the sth third-stage unit of the kth second-stage unit of the ith first-stage unit ($i=1,2,\ldots N$; $k=1,2,\ldots K_i$; $s=1,2,\ldots L$). We shall be concerned with the estimation of totals. The total T for a given stratum is given by

$$T = \sum_{i=1}^{N} \sum_{k=1}^{K_i} \sum_{s=1}^{L} Y_{iks} \qquad (1)$$

The linear unbiased estimate of T on the basis of a sample of n first-stage units, selected with probability proportional to the size of the units and without replacement from Frame A, and ℓ third-stage units selected with equal probability and without replacement from each of the second-stage units is

$$\hat{T}_A = \sum_{i=1}^{n} \frac{\sum_{k=1}^{K_i} \frac{L}{\ell} \sum_{s=1}^{\ell} Y_{iks}}{P_i} \qquad (2)$$

in which P_i is the probability of the first-stage unit i being included in a first-stage sample of n. In terms of the selection probabilities (which are chosen proportional to size of the FSU's), that is

$$p_i = K_i / \sum_{i=1}^{N} K_i, \quad i=1,2,\ldots N,$$

this inclusion probability when $n = 2$ is given by

$$P_i = p_i + \sum_{j(\neq i)=1}^{N} \frac{p_i p_j}{1-p_j} . \qquad (3)$$

The variance of $\hat{T}_A$ is given by

$$V(\hat{T}_A) = \sum_{i=1}^{N} Y_i^2 \left(\frac{1-P_i}{P_i}\right) + \sum_{i \neq j} Y_i Y_j \left(\frac{P_{ij}-P_i P_j}{P_i P_j}\right) + \sum_{i=1}^{N} \frac{1}{P_i} \sum_{k=1}^{K_i} L^2 \frac{S_{ik}^2}{\ell}\left(1-\frac{\ell}{L}\right) \qquad (4)$$

in which

P_{ij} = the probability of FSU's i and j being included in a first-stage sample of size n,

$$Y_i = \sum_{k=1}^{K_i} \sum_{s=1}^{L} y_{iks},$$

and

$$S_{ik}^2 = \sum_{s=1}^{L} (y_{iks} - y_{ik.})^2/(L-1)$$

where

$$y_{ik.} = \frac{1}{L} \sum_{s=1}^{L} y_{iks}$$

In the present study $N = 9$, $n = 2$, $L = 45$, $\ell = 2$. The essential theory underlying formulas (2) and (4) is due to Horvitz and Thompson (1951, 1952) and Narain (1951). When $n = 2$,

$$P_{ij} = p_i p_j \left(\frac{1}{1-p_j} + \frac{1}{1-p_i}\right).$$

The best linear unbiased estimate of T on the basis of a first-stage sample of size n from Frame B and third-stage samples of size ℓ selected from each of the second-stage units found in the n first-stage units, all selected at each stage with equal probability and without replacement is given by

$$\hat{T}_B = \frac{N}{n} \sum_{K=1}^{K} \frac{L}{\ell} \sum_{s=1}^{\ell} y_{iks} \qquad (5)$$

noting that $K_i = K = 32$ for all i in Frame B. The variance of $\hat{T}_B$, which can also be derived from (3) by putting $P_i = \frac{n}{N}$, $P_{ij} = \frac{n(n-1)}{N(N-1)}$ and $K_i = K$ and simplifying, is found to be

$$V(\hat{T}_B) = N^2 \left[\frac{S^2}{n}\left(1-\frac{n}{N}\right) + \frac{1}{nN} \sum_{i=1}^{N} \sum_{K=1}^{K} L^2 \frac{S_{ik}^2}{\ell}\left(1-\frac{\ell}{L}\right)\right] \qquad (6)$$

where $S^2 = \sum_{i=1}^{N} (Y_i - Y.)^2/(N-1)$ in which $Y. = \sum_{i=1}^{N} Y_i/N$.

Now the variance of the estimate having its basis in Frame A is $\sum_A V(T)$ and the corresponding variance for the estimate having its basis in Frame B is $\sum V(T_B)$ where the summation sign relates to summation over the five strata. Each of these expressions is a quadratic form in the underlying variates y_{iks}. The sign of the expression $\sum_A V(T) - \sum_B V(T)$ determines which of the two sampling procedures based on A or B is more efficient. Even for the case $n = 2$ it is not possible to determine the sign of the expression in the context of this study where $p_i = K_i/\sum K_i$. Generally the problem appears to be intractable.

The next approach is to make use of classical formulas for unbiased estimates of variance. This, too, has been avoided because the variability of the variance estimates might make the comparisons uncertain.

The approach which remains in such a situation is that of independent interpenetrating or replicated samples in the sense defined by Lahiri (1954). The technique consists of drawing two or more sets of samples from the same population using the same procedure of sampling for each set of samples. Sets of samples drawn in this manner are independent if and only if the sets of first-stage units selected are replaced after each drawing of a set is completed. This sometimes results in the same first-stage unit appearing in one or more sets.

For our study, we have drawn one hundred independent samples for each of the two sampling procedures under study.

The selection of first-stage units with

probability proportional to size of first-stage units and without replacement is very laborious if the usual method of cumulative totals is used. A simplified procedure of selecting first-stage units with unequal probability introduced by Lahiri was used.

One of us (Koop, 1960) has proved that the unbiased estimate of $V(\hat{T})$, the variance of an estimate T, from a set of m unbiased estimates

$$_1\hat{T},\ _2\hat{T},\ \cdots\ _m\hat{T}$$

each derived from independent replicated samples, whatever the underlying probability system and sample design, is given by

$$V(T) = \frac{\sum_{\alpha=1}^{m} (_\alpha T - \hat{T})^2}{m-1}$$

where $\hat{T} = \sum_\alpha \hat{T}/m.$ (7)

This result is used to compute the variances $\sum\hat{V}(\hat{T}_A)$ and $\sum\hat{V}(\hat{T}_B)$ on the basis of one hundred sets of independent estimates for each of the two sampling procedures. The variances of course are computed stratum by stratum and then added as indicated by the respective formulas.

Regarding the estimates for strata it may be noted that $\hat{T}_B$ is self weighting and is given by $\frac{9 \times 45}{4} \sum_i \sum_k \sum_s y_{iks}$, whereas $\hat{T}_A$ is not so and is given by $22.5 \sum_{i=1}^{2} \frac{1}{P_i} \sum_k \sum_s y_{iks}$. Each P_i is computed using formula (3).

6. Data Used for Investigation

The following four characteristics studied in the Nashville Morbidity Survey are selected for the purpose of comparing the frames:

(1) Number of people in households
(2) Number of deaths reported
(3) Number of employed individuals
(4) Number of households bothered by smog.

In the N.M.S. the selected sampling units were designated by five digit numbers. The first two digits were for the planning unit, the next two referred to the stratum and the last represented the sampling unit. The data is available on punched cards and is shown in skeleton form in Table 4.

Table 4. Data for sampling units selected in N.M.S.

Sampling Unit	Characteristics (1)	(2)	(3)	(4)
12 01 1	5	0	1	0
12 01 2	4	0	3	0
==				
12 19 1	2	0	2	0
12 19 2	4	1	0	0
13 01 1	2	0	2	0
13 01 2	3	0	1	0
==				
13 22 1	4	0	1	0
13 22 2	5	0	1	0
--				

For the present study, the above data is used as follows:

Frame A: The cards are first sorted and then grouped for each of the 45 first-stage units. The totals for each of the four characteristics are found by running the cards in I.B.M. 407 Tabulator. These totals for each of the 45 FSU's are used in the sampling design based on Frame A. The card numbers running from 12 01 1 to 12 19 2 fall in the first FSU. The allocation of the card numbers to the different FSU's and their totals for each of the four characteristics are given elsewhere (Khosla, 1961).

Frame B: We first make the necessary divisions of the cards to correspond to the new FSU's. For example cards running from 12 01 1 to 13 13 2 are put in the first FSU and from 13 14 1 to 14 23 2 in the second FSU. These totals are also shown in the above reference.

The totals from Frame B can be used as they are for the computation of stratum variances but the totals for each FSU from Frame A have to be multiplied by the relevant factor $\frac{1}{P_i}$ as indicated in the last paragraph of the previous section.

7. Comparison of Precision and Costs

The estimates for 4 characteristics studied are given in Table 5.

In three of the four characteristics under study the estimate of variance based on Frame A is greater than the corresponding estimate of variance based on Frame B. The relative precision of the sampling procedure based on Frame B compared to the sampling procedure based on Frame A is given by the ratio of the estimate of variance for Frame A to the corresponding estimate of variance for Frame B. The increase in the relative precision of Frame B expressed in percentage is found to be 27.65, 32.90, and 20.97 for the characteristics (1) number of people in households, (3) number of employed individuals and (4) the number of households bothered by smog, respectively. These results are shown in Table 6. In characteristic (2), number of deaths, there is a decrease of 0.62 per cent in the relative precision of Frame B.

There are 36^5 distinct samples possible for each of the two sampling procedures used. We have selected only a set of a hundred samples for each. It should be noted that any inference based on a sample study is subject to the usual uncertainties of sampling.

The costs involved in the two schemes are considered under the following headings:

(1) cost of constructing frames,
(2) interviewing costs in a sample under the two procedures,
(3) computing costs.

Table 5. Estimates of totals for entire population*

	From Frame A	From Frame B
(1) Number of people in households	200299	201673
(2) Number of deaths reported	7725	7805
(3) Number of employed individuals	70210	70878
(4) Number of households bothered by smog	14285	14457

*Each of these estimates are the means of the hundred independent estimates based on the relevant formula (2) or (4).

Table 6. Estimates of variances for each procedure

	$\sum_{h=1}^{5} \hat{V}(\hat{T})_{h\,A}$ Frame A	$\sum_{h=1}^{5} \hat{V}(\hat{T})_{h\,B}$ Frame B	Relative precision of Frame B (per cent)
(1) Number of people in households	89857314.5	70394481.1	127.65
(2) Number of deaths reported	776405.25	781252.06	99.38
(3) Number of employed individuals	10721893.6	8067576.8	132.90
(4) Number of households bothered by smog	4224824.8	3492453.1	120.97

a/ A multiplying factor of 1/100 applies to variances of estimates given in Table 5 since each of these estimates is based on 100 sets of independent estimates.

It was difficult to evaluate the comparative costs of constructing the two frames as the materials used for our study were already available. It can, however, be stated that it is possible to construct an area frame of equal sized first-stage units from the different materials available for area sampling work in this country. The additional material expense for constructing Frame B would be in the form of extra maps and attendant materials necessary for redefining areas of equal sized first-stage units.

Although we have ensured that the total sets of samples under each scheme is 100, there is no possible way to make the number of sampling units equal for each sample. The total number of sampling units in a sample under the two sampling schemes are

$$2\sum_{h=1}^{5}\sum_{i=1}^{2} K_{hi}$$

for Frame A and 640 for Frame B. The expected number of sampling units based on a sample from Frame A is 804.8. The number of sampling units in our set of a hundred samples based on Frame A was found to vary from 662 to 948 with a mean of 801.5. For the sampling system considered, the cost of enumeration using Frame A can be expected to be higher.

The computing time with Frame B was roughly thirty man-hours. This included all operations of computations manually performed with the Monromatic machine and without any recourse to I.B.M. Compared to this, seventy man-hours were necessary for the initial set up alone for Frame A and in addition I.B.M. was used to calculate estimates and their variances. On the basis of this experience, it seems that the computing time involved with Frame A (unequal probability sampling) with always be higher than with Frame B (equal probability sampling), whenever n the first-stage sample size from each stratum is greater than one.

With Frame B the selection of more than two FSU's from each stratum can be made with a proportionate increase in the computing cost. Compared to this, the cost of computation with Frame A will increase at a much faster rate as we progressively increase the number of first-stage units in the sample.

Our study at least shows that sampling with equal sized first-stage units deserves more attention.

References

Finkner, A. L., Monroe, J., and Fleischer, J. (1960) The Nashville Morbidity Survey, Institute of Statistics Mimeo. Series No. 252. N. C. State College, Raleigh, N. C.

Horvitz, D. G. and Thompson, D. J. (1951) A generalization of sampling without replacement from a finite universe. Abstract. Ann. Math. Stat., 22 (2):315

_____ (1952) Ibid. J. Amer. Statist. Assoc. 47:663-685.

Khosla, T. (1961) Comparison of two frames with equal and unequal sized first-stage units. Unpublished M. S. Thesis, N. C. State College Library.

Koop, J. C. (1960) On theoretical questions underlying the technique of replicated or interpenetrating samples. 1960 Proceedings of the Social Statistics Section. Amer. Statist. Assoc., 196-205.

Lahiri, D. B. (1954) National Sample Survey: Some Aspects of Sample Design. Sankhya, 14: 268-316.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. J. Ind. Soc. Agric. Statist., 3 (2):169-174.

Vaivanijkul, N. (1961) A comparison of respondents and non-respondents in a sample survey. Unpublished M. S. Thesis, N. C. State College Library.

III

PREDICTION IN DELINQUENCY AND PAROLE

Chairman, Clarence Schrag, University of Washington

Page

Development and Operational Use of Prediction Methods in Correctional Work - Don M. Gottfredson, Institute for the Study of Crime and Delinquency and Robert F. Beverly, California Youth Authority..... 54

Behavior Factors in Gang Delinquency - James F. Short, Jr. and Ray A. Tennyson, Washington State University and Kenneth I. Howard, University of Chicago..... 62

Theory Construction and Prediction in Juvenile Delinquency - Leroy C. Gould and Clarence Schrag, University of Washington... 68

DEVELOPMENT AND OPERATIONAL USE OF PREDICTION METHODS IN CORRECTIONAL WORK[1]

Don M. Gottfredson, Ph.D., Institute for the Study of Crime and Delinquency
AND
Robert F. Beverly, California Youth Authority

This paper summarizes the development and validation of several devices purporting to predict behavior of California prisoners and Youth Authority wards following their release under parole supervision. The utility of these devices for studies of treatment effects will be described. Some current attempts to improve prediction will be reported, and two practical operational uses will be discussed.

DEVELOPMENT OF THE PREDICTION DEVICES

A number of investigators have used a variety of methods in attempting development of useful parole prediction measures. These are well summarized by Mannheim and Wilkins (23), who, while studying the British Borstal System used a multiple linear regression method.

Three separate prediction devices, called "Base Expectancy Measures", were developed from sample studies of California adult male and female prison populations, and of the male California Youth Authority population. Adult male subjects were 873 men, selected (under a procedure assumed to approximate random selection) from all who were released to California parole supervision in 1956 (11). Adult women studied were 695 persons released to California parole supervision between July 1, 1955 and June 30, 1958 (9), and Youth Authority wards were all 11,435 boys released to California parole supervision in 1956 through 1958.

In each of these three studies, a multiple linear regression analysis was completed, with a dichotomous parole adjustment criterion (scored 0 or 1) as the dependent variable. The parole adjustment criterion definitions may be summarized as follows:

For the adult male subjects, "favorable parole adjustment" was considered to have occurred if and only if the subject was not classified as a parole violator by reason of return to prison, sentence to 90 days or more in jail, or identification as a parole violator at large (absconder) before two years following release.

For adult women, "favorable adjustment" was said to have occurred if and only if the subject was not returned to prison from parole over a two year follow-up period.

For Youth Authority wards, "favorable adjustment" was identified with absence of revocation of parole or with discharge from a suspended parole status within 15 months of parole exposure.

The forms shown in Figures 1, 3, and 5 summarize the prediction methods.

In each case, the resulting prediction equations employ predictor items selected from a larger pool of independent variables selected for study. Following calculation of a correlation matrix and the regression coefficients, the proportion of variance in the dependent variable attributable to inclusion of each independent variable was calculated. By this means, predictor items failing to add appreciably to R^2 (the coefficient of determination) were dropped.[1,2] For definitions of predictor items see (3) (8) and (9).

---

1. These opinions of the authors do not necessarily reflect the views or endorsement of the Institute For The Study of Crime and Delinquency or The Youth And Adult Corrections Agency. Portions of the research reported were conducted under funds provided by the National Institute of Mental Health (Project OM 823).

1. As illustration, the following procedure was followed in the case of construction of the Base Expectancy formula for women (CDC-BE-CIW 62A): After calculation of the matrix of Pearson Product Moment Correlation Coefficients in the usual way, the Gaussian Multipliers were computed following the procedure given by Ostle (24). These were then used to calculate the regression coefficients, the sum of squares attributable to regression, and the sum of squares of deviations from regression. The ratio of these two sums of squares equals R^2. The sum of squares for each additional regression term (and the proportion of variance added by its inclusion) was calculated following the procedure given by Bennett and Franklin (2).

2. A number of individuals and institutions provided help or consultation in conducting these analyses. Appreciation is expressed to Leslie Wilkins, Deputy Director, Research Unit, British Home Office; Professor Robert V. Oakford, Department of Industrial Engineering, Stanford University, and the Stanford Computation Center; William P. Anderson, Chief of User Services, Western Data Processing Center, University of California at Los Angeles; Gordon Rowe, Extension Economist, University of California, Berkeley, and the University of California Computation Center.

CALCULATION OF BASE EXPECTANCY RAW SCORES
MEN (FORM CDC BE 61B)

	IF	ADD
A.	ARREST FREE 5 OR MORE YEARS	16 ___
	NO HISTORY OF ANY OPIATE USE	13 ___
	NO FAMILY CRIMINAL RECORD	8 ___
	NOT CHECKS OR BURGLARY	13 ___
B.	AGE AT COMMITMENT TIMES .6	___
	21 IS ADDED FOR ALL CASES	21
C.	SUBTOTAL: A+B	
D.	ALIASES: -3 TIMES NUMBER	- ___
E.	PRIOR INCARCERATIONS: -5 TIMES NUMBER	- ___
F.	SUBTOTAL: D+E	-
G.	B.E. RAW SCORE: SUBTRACT F FROM C	

FIGURE 1
CALCULATION OF BASE EXPECTANCY SCORES (MEN)

CALCULATION OF BASE EXPECTANCY RAW SCORES
CALIFORNIA INSTITUTION FOR WOMEN
(FORM CDC-BE-CIW 62A)

A.	ARREST FREE YEARS	___ A
B.	81 IS ADDED FOR ALL CASES	81 B
C.	SUBTOTAL: A+B	C
D.	HEROIN: ANY HISTORY OF USE -24	- ___ D
E.	ALCOHOL: ABSTAINER -0 AT OFFENSE -12 USER -4 "ALCOHOLIC"-16 DIFFICULTY -8	- ___ E
F.	ALIASES: -4 TIMES NUMBER	- ___ F
G.	PROBATION OR PAROLE VIOLATIONS: -3 TIMES NUMBER	- ___ G
H.	PRIOR ARRESTS: - NUMBER	- ___ H
I.	SUBTOTAL: D THROUGH H	-
J.	B.E. RAW SCORE: SUBTRACT I FROM C	

FIGURE 3
CALCULATION OF BASE EXPECTANCY SCORES (WOMEN)

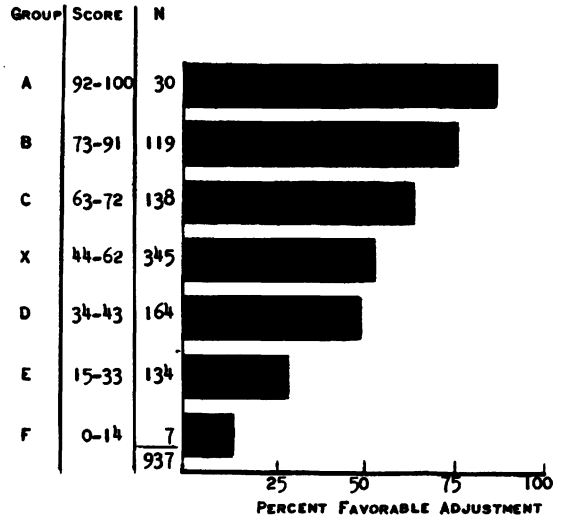


FIGURE 2
RELATIONSHIP OF BASE EXPECTANCY SCORES (MEN) TO PAROLE ADJUSTMENT IN VALIDATION SAMPLE

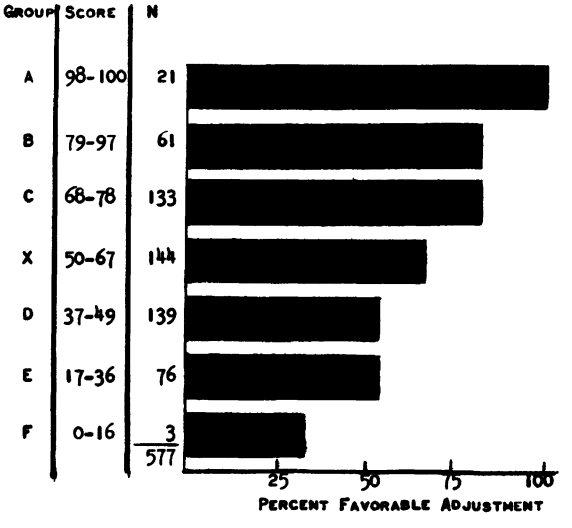


FIGURE 4
RELATIONSHIP OF BASE EXPECTANCY SCORES (WOMEN) TO PAROLE ADJUSTMENT IN VALIDATION SAMPLE

CALCULATION OF BASE EXPECTANCY SCORES
CALIFORNIA YOUTH AUTHORITY

IF	ADD
A. CRIME VS. PERSONS	9
CRIME VS. PROPERTY	0
B. COUNTY GROUP ONE	10
TWO	5
THREE	0
C. AGE AT FIRST ADMISSION	
16 OR OLDER	12
15	6
14 OR YOUNGER	0
D. COURT OF MOST RECENT COMMITMENT	
JUVENILE	6
ADULT	0
E. AGE AT RELEASE	
18 OR OLDER	10
17	4
16 OR YOUNGER	0
F. ADMISSION STATUS	
1ST ADMISSION OR RETURN WITH NEW COMMITMENT	8
RETURN WITHOUT NEW COMMITMENT	0
G. PRIOR RECORD	
NO PRIOR COMMITMENT AND NO MORE THAN TWO PRIOR DELINQUENT CONTACTS	13
NO PRIOR COMMITMENT AND THREE OR MORE PRIOR DELINQUENT CONTACTS	5
PRIOR COMMITMENTS	0
TOTAL POSSIBLE SCORE	68

FIGURE 5
CALCULATION OF BASE EXPECTANCY SCORES
(YOUTH AUTHORITY)

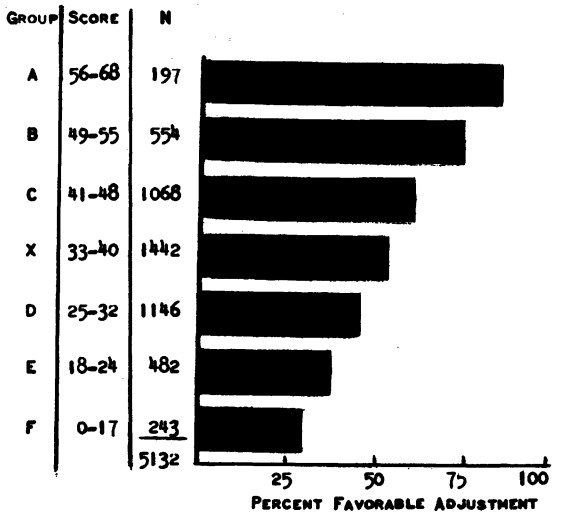


FIGURE 6
RELATIONSHIP OF BASE EXPECTANCY SCORES
(YOUTH AUTHORITY) TO PAROLE ADJUSTMENT
IN VALIDATION SAMPLE

TABLE I
BASE EXPECTANCY SCORES FOR MEN (FORM CDC BE 61B) AND NUMBER AND PERCENT WITH FAVORABLE PAROLE ADJUSTMENT

BASE EXPECTANCY SCORE	NUMBER		PERCENT	
	FAVORABLE	UNFAVORABLE	FAVORABLE	TOTAL
92-100	26	4	87	30
73-91	90	29	76	119
63-72	89	49	64	138
44-62	183	162	53	345
34-43	81	83	49	164
15-33	39	95	29	134
0-14	1	6	14	7
TOTAL	509	428	54	937

	FAVORABLE	UNFAVORABLE	TOTAL
MEAN	57.36	46.79	52.53
S.D.	17.80	17.32	18.34

DIFFERENCE BETWEEN MEANS = 10.57
P < .0001
BISERIAL CORRELATION COEFFICIENT .36
POINT BISERIAL CORRELATION COEF. .29

TABLE II
BASE EXPECTANCY SCORES FOR WOMEN (FORM CDC BE CIW 62A) AND NUMBER AND PERCENT WITH FAVORABLE PAROLE ADJUSTMENT

RAW SCORE	NUMBER		PERCENT	
	FAVORABLE	UNFAVORABLE	FAVORABLE	TOTAL
98-100	21	0	100	21
79-97	50	11	82	61
68-78	109	24	82	133
50-67	96	48	67	144
37-49	75	64	54	139
17-36	41	35	54	76
0-16	1	2	33	3
TOTAL	393	184	68	577

	FAVORABLE	UNFAVORABLE	TOTAL
MEAN	62.06	50.35	58.32
S.D.	19.36	16.36	19.23

DIFFERENCE BETWEEN MEANS = 11.71
P < .0001
BISERIAL CORRELATION COEFFICIENT .37
POINT BISERIAL CORRELATION COEF. .28

TABLE III

BASE EXPECTANCY SCORES FOR BOYS (CALIFORNIA YOUTH AUTHORITY) AND NUMBER AND PERCENT WITH FAVORABLE PAROLE ADJUSTMENT

Base Expectancy Score	Number		Percent	
	Favorable	Unfavorable	Favorable	Total
56-68	161	36	82	197
48-55	400	154	72	554
41-48	634	434	59	1068
33-40	751	691	52	1442
25-32	508	638	44	1146
18-24	178	304	37	482
0-17	70	173	29	243
Total	2702	2430	53	5132

	Favorable	Unfavorable	Total
Mean	55.20	43.16	51.18
S.D.	22.55	20.29	22.53
Difference between means		=	12.04
		P <	.0001
Biserial Correlation Coefficient			.33
Point Biserial Correlation Coef.			.25

VALIDATION STUDIES

Results of application of these measures to validation samples are summarized in Figures 2, 4, and 6 and Tables I, II, and III.[1] Adult men and women subjects, released to California parole supervision, were 937 men released in 1956 and 577 women released between July 1, 1958 and June 30, 1960. The men were selected from all releasees during that period (under a procedure assumed to approximate random selection) while the women were all who were released. Youth Authority wards were all 5,132 boys released in 1960. In each case "favorable parole adjustment" is a decreasing function of Base Expectancy Scores.

These or similar prediction methods have demonstrated validity in samples released in different years, in cross validation samples (3) (9) (10) (11) (12) (13), in samples released to different geographical areas and at different seasons of the year (20), and in samples released from different institutions (16).

1. In Figures 2 and 4 and in Tables I and II, Base Expectancy scores are arbitrarily grouped in terms of the standard deviations; i.e., Group F$\leq -2\sigma$, E= -1σ to -2σ, D= $-.5\sigma$ to -1σ, x= $-.5\sigma$ to $.5\sigma$, C= $.5\sigma$ to 1σ, B= 1σ to 2σ, and A$\geq 2\sigma$. Base Expectancy Scores in Figure 6 and Table III are grouped by dividing all possible scores by seven.

These devices have been called "Base Expectancies" because they provide a base for further research by quantifying our expectations. They do this by summarizing our experience. All Correctional workers know that parolees are not equally likely to violate the conditions of their parole. A "Base Expectancy" is a statement of the expected parole success rate for a given group; and this statement is made on the basis of past experience with other similar groups.

THE ROLE OF BASE EXPECTANCIES IN STUDYING TREATMENT EFFECTS

What kinds of treatment help what kinds of inmates in attaining favorable parole adjustment?

To test the large number of null hypotheses generated by this general question, experimental arrangements according to classical experimental design are needed. This is the best test of treatment effects, particularly if the experiment is replicated.

But a number of considerations argue for consideration of an alternative approach. Besides problems associated with representative sampling of subjects and treatments and with attempts to control all variables but one,[1] there is the question of which treatments to study by experimental designs. There are so many differing programs in correctional work, all actively supported by ardent advocates, it is not administratively feasible to test them all through this kind of research design.

A study of decisions, as proposed by Wilkins (25)(26), is an alternative with promise of provision of helpful information to the correctional decision-maker. Briefly, the outline for this approach to studying treatment effects is as follows:

> Base Expectancies, like those discussed above, defined as the probability of favorable outcome, are needed for each person before assigned to treatment. When persons are assigned specific kinds of treatments, then is the actual outcome more or less favorable than expected? We wish to find treatments that improve the chances of favorable outcomes, and we will therefore be pleased if the Base Expectancy is made invalid by treatment helpful to the inmate.

1. See Brunswick, E. Systematic and Representative Design of Psychological Experiments. Berkeley: Univ. of Calif. Press, 1949 (6).

> If the outcome following treatment can be predicted not only before treatment but regardless of treatment, it is very hard to argue that this treatment makes any difference with respect to this outcome.[1] It may be, however, that persons assigned to one treatment tend to succeed (or fail) significantly more often than expected from analysis of the "kinds of risks" assigned the treatment. If the validity of the expectancy has been established on other groups such differences must be due to either treatment or factors associated with treatment, or both.
>
> Further research, using experimental designs, can identify the source of the difference. Meanwhile, correctional decision-makers can be given knowledge of the relationship (or lack of it) between treatment and outcome.

A study of all 8,723 boys committed to the California Youth Authority by juvenile courts and later released to California parole supervision in 1956, 1957, and 1958 illustrates the procedure suggested above (4). Subjects were classified by institution of release, year of release, and a Base Expectancy categorization. Chi square tests were employed to test the hypothesis that parole adjustment is independent of institution of release. The classification by Base Expectancy scores was assumed to control for selection factors known before institution assignment.

The null hypothesis was supported in the case of 5 of the 8 institutions studied.[2] This was interpreted as indicating that the observed differences in violation rates among releasees of these institutions were satisfactorily explained by the "type of inmate" assigned the institution. Of the three remaining institutions, releasees of the reception center clinics and forestry camps did significantly better, and releasees of one institution did significantly worse, than expected. These differences were attributed to unknown selection factors and/or treatment, the effects of which could not be completely isolated on the basis of these data. Further research is needed to identify the bases for the differences found.

1. Aside from possible "masking effects"; recent research demonstrates this possibility (1) (17) (18) (19).

2. Data and analyses supporting this contention are contained in Beverly, R. F., "An Analysis of Parole Performance by Institution of Release," Research Report No. 22, Sacramento: Dept. of the Youth Authority, April 1961 (4).

THE CLINICAL VS. STATISTICAL PREDICTION CONTROVERSY

Twenty years ago a psychologist observed "the statistician and the case study investigator can make mutual gains if they will quit quarreling with each other and begin borrowing from each other" (22). It is time we heeded this advice. As pointed out by de Groot (7), there is more to be gained through efforts to improve statistical prediction via clinical prediction than through continuation of arguments or comparisons of predictive accuracy of the two approaches. Further, subjective judgment and statistical prediction should be cooperative--rather than competitive--functions in a single decision process (11).

Since prediction of outcomes is always uncertain, at best we can deal with probabilities. Any prediction strategy will misclassify some persons. The statistician (correctly) regards this as "error" and he usually can state with some confidence its probable limits. The clinician, however, may (correctly) regard this as individuality or uniqueness (of personality or situation) beyond that measured by the statistical prediction method.

This suggests a way to improve our predictive ability and hence our decisions. Give the decision-maker the statistical prediction device. He can add any information believed relevant and arrive at his best predictive statement. If it can be demonstrated that prediction is improved by this process, we can then attempt to identify the information used. Then the statistician can include it in a new, more adequate tool.

One recent study attempted to not only compare but also to combine subjective predictions with the Base Expectancy score (13). The validity of the objective method (Figure 1) with Base Expectancy scores obtained by clerks was compared (for the same 283 subjects) with subjective ratings (based on pre-release interviews) by an associate superintendent and also with his ratings of clinical council (psychiatric) reports. The two subjective prediction ratings and the Base Expectancy scores all were related significantly to parole outcomes (in the expected direction), though the clerks were much better predictors.

When the subjective and objective ratings were combined, however, the subjective ratings added nothing to the predictive accuracy of the simple check list. Correlation coefficients describing the relationship between prediction scores and the parole outcome were .20 for the associate superintendent ratings, .21 for the clinical council ratings, and .48 for the Base Expectancy scores. Multiple correlation of all combinations of the three measures failed to increase the value or R beyond .48. But the fact that the subjective ratings in this study did not improve prediction does not mean this may not occur if other methods are used or

other persons make them. The method provides a way to test and utilize helpful subjective judgments.

Base Expectancy scores are now routinely collected in the Reception-Guidance Centers of the California Department of Corrections. In addition to obtaining the information needed, clinicians make a subjective estimate of expected parole outcome. These data can lead to improvement of prediction along the lines suggested above.

SOME FURTHER EFFORTS TOWARD IMPROVED PREDICTION

Attempts to identify additional predictive information include, in the California Youth Authority, completion of a Home Visit Research Schedule during one of the agency's earliest contacts with each case (5). The schedule was developed to tap a number of areas particularly accessable during the parole agent's initial home visit. When 900 releasees were later followed on parole, a number of items differentiated favorable and unfavorable parole adjustment (significant at the five percent level of confidence as determined by Chi square tests).[1]

For example, boys later classified as attaining a favorable parole adjustment were, at the initial home visit, reported less often to have had histories of truancy or other "serious or persistent" school misbehavior from ages eight to fifteen. Less often they were rated as markedly disliking school. Those later classed as parole violators were less often involved with other persons in the offense leading to commitment, and more often there was a record of broken homes. Boys who had lived three or more years at their current address tended to fall in the favorable group.

These items have been tested with a validation sample of 1035 wards. A number of the items, (including those mentioned above) held up. Most of the remaining items showed a trend consistent with the results of the first sample.

When the two samples were combined, twenty-five items were significantly related to parole violation (six at the .05 level and nineteen at the .01 level), and showed consistent and substantial trends in the two samples.

Once Initial Home Visit information on all male first admissions to the California Youth Authority has been obtained for a full release cohort, these items along with those already in use will be used as the basis for a new base expectancy equation which will hopefully improve predictive efficiency.

1. This sample consists of the first 900 boys released following initiation of the home visit Research Schedule; therefore the sample is biased by restriction to those serving relatively short institution terms, and these results are viewed as suggestive only.

Additional research in progress obtains measures of personality just before release, attempts to identify program participation and institution adjustment indices with hypothesized relationships to parole performance, and seeks improvement of outcome measures used (14) (15).

OPERATIONAL USES OF BASE EXPECTANCIES

The devices reported provide tools useful in both individual case decisions and agency program decisions (15). If their limitations are understood, then they can be used to advantage when combined with other relevant information about the individual.

A recent screening of the entire California prison population, first by Base Expectancy scores, then by clinical criteria, has been accomplished (8). The result was a small group of men referred for parole consideration at a date earlier than originally scheduled. Also, each man and woman appearing before the parole authority now has a Base Expectancy score in addition to the information otherwise collected for assistance in the release decisions.

Minimal supervision case loads of both male and female parolees have been established for persons classed as having a high probability of successful parole completion; and experience has demonstrated that these cases may be given less supervision with no increase in the parole violation rate (21). This enables parole workers to deploy their forces from areas where help is less needed to concentrate efforts where it might be more helpful.

POTENTIAL USE OF PREDICTION METHODS IN OTHER AREAS OF CORRECTIONAL WORK

While this paper, and most other work involving prediction methods in corrections, has focused on the problem of parole behavior prediction, the same method may profitably be employed in efforts to increase our predictive ability in other areas.

For example, a study of 864 adult men released to California parole supervision in 1956 was completed to explore problems of prediction of time to be served (under indeterminate sentencing) in prison and on parole. Using eight predictor variables including the legal offense, a rating of its "severity", age, and prior record, a multiple correlation coefficient of .64 was obtained with the number of months served in the institution. Four of these items in combination correlated .60 with number of months granted on parole. These Time Served Expectancy Measures have not yet been tested in validation samples, but they suggest the feasibility of the approach. Such measures can be helpful in program planning, in description of the paroling decision, and in program studies.

Whether either multiple regression or discriminant function methods can be helpful in such areas as prediction of institution adjustment, of response to psychotherapy, of escape, or of specific kinds of delinquency behavior after release remains to be investigated.

REFERENCES

1. Adams, S., "Interaction Between Individual Interview Therapy and Treatment Amenability in Older Youth Authority Wards," Board of Corrections Monograph No. 2, Sacramento: Printing Division, Documents Section, July, 1961.

2. Bennett, C. A., and Franklin, N. L., Statistical Analysis in Chemistry and the Chemical Industry; Appendix 6A, John Wiley and Sons, 1954.

3. Beverly, R. F., "A Method of Determination of Base Expectancies for Use in the Assessment of Effectiveness of Correctional Treatment," Research Report No. 3, Sacramento: Department of the Youth Authority, Feb., 1959, unpublished manuscript.

4. Beverly, R. F., "An Analysis of Parole Performance by Institution of Release," Research Report No. 22, Sacramento: Department of the Youth Authority, April, 1961.

5. Beverly, R. F., "The Initial Home Visit Research Schedule," Research Newsletter, No. 5, Vacaville, California Medical Facility, California Department of Corrections, October, 1960.

6. Brunswik, E., Systematic and Representative Design of Psychological Experiments, Berkeley: Univ. of California Press, 1949.

7. de Groot, A. D., "Statistical via Clinical Prediction," invited address, Western Psychological Association, San Jose, April, 1960.

8. Department of Corrections Administrative Bulletin 61/115, "Provision of Base Expectancy Information to Adult Authority," Sacramento: Department of Corrections, December, 1961.

9. Gottfredson, D. M., Ballard, K. B., and Bonds, J. A., Base Expectancy (Form CDC-BE CIW 62A) California Institution for Women, Sacramento: Institute for the Study of Crime and Delinquency and Research Division, California Department of Corrections, July, 1962 (in draft).

10. Gottfredson, D. M., and Bonds, J. A., A Manual for Intake Base Expectancy Scoring (Form CDC-BE 61A), Research Division, California Department of Corrections, Sacramento, California, April 1, 1961.

11. Gottfredson, D. M., Bonds, J. A., and Grant, J. D., "La combinazione della previsione clinica e di quella statistica nelle decisioni penitenziarie", Quanderni di Criminologia Clinica, n. 1, Roma: Tripografia Delle Mantellate, Gennaio, Marzo, 1962.

12. Gottfredson, D. M., "A Shorthand Method for Base Expectancy Scoring (Form CDC-BE 61C)", Research Report No. 4, Sacramento: Research Division, California Department of Corrections, Dec., 1961.

13. Gottfredson, D. M., "Comparing and Combining Subjective and Objective Parole Predictions," Research Newsletter, No. 3, Vacaville: California Medical Facility, September-December, 1961.

14. Gottfredson, D. M., "One Way of Studying a Black Box", paper presented to the Ninth National Institute on Crime and Delinquency, Seattle, July, 1962.

15. Gottfredson, D. M., "Studying Correctional Decisions," paper presented to the Ninth National Institute on Crime and Delinquency, Seattle, July, 1962.

16. Gottfredson, D. M., "The Role of Base Expectancies in the Study of Treatments," paper presented as part of the Symposium on Methods for the Study of Effectiveness of Treatment, Western Psychological Association meeting, April, 1959, San Diego, California.

17. Grant, J. D., and Grant, Marguérite Q., "A Group Dynamics Approach to the Treatment of Nonconformists in the Navy," Annals Amer. Acad. Pol. and Soc. Sci., 1959, 322, 126 - 35.

18. Grant, J. D., "The Case for Studying Interactions Between Kinds of Treatment and Kind of Delinquents," paper presented at Conference on Research and Potential Application of Research in Probation, Parole, Delinquency Prediction, October, 1960, New York, sponsored by Research Center, N. Y. School of Social Work, Columbia University and Citizens Committee for Children of New York, Inc.

19. Gunderson, E. K., and Ballard, K. B., "Changes in Nonconformist Attitudes Induced by Closed Living Groups," Research Newsletter, No. 5, Vacaville: California Medical Facility, California Department of Corrections, October, 1960.

20. Havel, J., and Sulka, E., "Special Intensive Parole Unit, Phase Three," Research Report No. 3, Sacramento: Research Division, California Department of Corrections, March, 1962.

21. Havel, J., "Special Intensive Parole Unit, Phase Four," personal communication.

22. Horst, Paul, Ed., The Prediction of Personal Adjustment, Social Science Research Council, New York, 1941.

23. Mannheim, H., and Wilkins, L. T., Prediction Methods in Relation to Borstal Training, 1955, London: Her Majesty's Stationery Office.

24. Ostle, B., Statistics in Research, The Iowa State College Press, 1954.

25. Wilkins, L. T., "Symposium on Predictive Methods in the Treatment of Delinquency, I," Brit., J. Delinq., 1955, 6, 82 - 98.

26. Wilkins, L., "What is Prediction and is it Necessary in Evaluating Treatment?" Research and Potential Application of Research in Probation, Parole and Delinquency Prediction, New York: Citizens' Committee for Children of New York, Inc., and Research Center, New York School of Social Work, Columbia University, July, 1961.

BEHAVIOR FACTORS IN GANG DELINQUENCY

James F. Short, Jr. and Ray A. Tennyson, Washington State University
and Kenneth I. Howard, University of Chicago

In recent years there has been increasing theoretical concentration on the specification of types or patterns of delinquency adaptation. It has become clear that the notion of a "general delinquency" adaptation has limited descriptive and explanatory promise. However, the question as to the existence and description of specialized delinquent adaptations is in large measure an empirical one.

In this paper we shall describe and present preliminary findings of a comprehensive empirical study which sought answers to the following questions:

a) To what extent does a "general delinquency" dimension account for *in situ* street corner behavior?

b) What evidence is there for specialized delinquent adaptations?

c) If there are specialized adaptations, what are their characteristics?

The Research Setting

Data were collected from 598 members of 16 "delinquent gangs" assigned detached workers by the Program for Detached Workers of the YMCA of Metropolitan Chicago. The gangs ranged in size from 16 to 68 members, on the basis of workers' judgments concerning who should and who should not be considered members. Initially gangs were selected by the YMCA on the basis of their generally troublesome character to the community, as judged by police complaints and the reports of welfare agencies, and by field investigations of the Detached Worker staff. Later, in collaboration with the University of Chicago research program, gangs were selected in an attempt to fulfill requirements of a research design which sought to study gangs representative of major "delinquent subcultures" discussed in the literature.[1]

The search for gangs concentrated on the "criminal" and "drug use" types, principally because conflict-oriented gangs, particularly among Negroes were abundant, and for theoretical reasons which are discussed elsewhere.[2] Finding gangs whose primary activities and norms were oriented around drug use and rational, systematic, economically motivated criminal activity proved to be a major problem. Much marihuana smoking was found in several groups and heroin and "pills" were experimented with by many boys, always in concert with other boys, it seemed, but not as a major activity or normative emphasis of the groups under observation. It was not until the fall of 1960, after more than a year of extensive inquiries among police and local community adults, coupled with field investigations by detached workers and our own research personnel, that a drug-oriented group was located. No criminal group was located on the basis of these unsystematic and impressionistic criteria, though much criminal activity, apparently economically motivated, was found in nearly all groups.

The Data

Dissatisfaction with official statistics as a basis for measuring delinquent behavior has led in recent years to experimentation with a variety of methods for obtaining better measures of such behavior than are provided by official records.[3] Careful consideration of alternative methods of generating data concerning the behavior of the gang boys under observation led to the rejection of most of the methods reported in the literature, and the development of a system of ratings by the detached workers as at once the most feasible and reliable method of generating such data. The availability of detached workers as intimate observers of the boys, particularly in the gang setting, offered a rare opportunity to gain more complete and objective insights into the behavior of these boys than could be provided by any other method. Weekly interviews with the workers convinced us that they shared intimately in the on-going life of the gang and of many of its members.[4] Such behavioral information as was not known by direct contact with a boy could usually be inferred from conversations among the boys or directly with the worker in the endless "bull sessions" on the street corner. The nature of these contacts seemed particularly conducive to objective reporting of the type of behavior we were interested in measuring, i.e., street corner behavior. This behavior could be described in concrete terms and it did not require abstract conceptualization by our informants. Some of the behavior was directly observable by the workers. Information as to the remainder was readily accessible to them.

What we needed was a "baseline" of information on the behavior of our population, information which could be statistically manipulated and related to other variables we were studying. Toward this end a list of 69 behavioral categories representing a wide range of delinquent and nondelinquent behaviors was drawn up. The list included behavior observed in other studies of street corner life, by police as well as researchers and other observers. Each behavior was carefully defined for the workers and their ratings were discussed in detail in order to insure standardization and accuracy of reporting. Because we were interested more in incidence of behaviors and their relative frequency, rather than absolute frequencies, and because workers' observations could not extend to actual counts of behavior, we asked only that each boy be rated in terms of whether or not he had engaged in a

particular type of behavior only a few times or more often than this. Workers were asked to report only after they had been in contact with a group for at least six months, to restrict their reporting to boys whom they and the group recognized as group members, and to limit their reports to information in which they were very confident. An item-by-item review of their reports was undertaken by the research staff in order to insure the most reliable reporting possible.

It should be noted that each gang member was rated by only one rater. Consequently, no reliability checks were available. We had to assume that obtained differences between gangs were due to differences in behavior rather than differences in rater characteristics. In view of the procedures described above concerning the training of the raters, the interview review of all ratings and the relatively objective character of the behaviors rated, we have some confidence in this assumption.

To facilitate statistical manipulation of the data, the 69 behaviors were reduced by combining items into 37 items on the basis of assumed similarity of item content. Product moment correlations were calculated between all 37 items. The generally positive character of the resulting correlation matrix suggested that involvement in delinquency is to some extent a global phenomenon, the dimensions of which might lie along a continuum of more or less delinquency involvement, regardless of the specific offenses committed. The considerable range within our population in the incidence of different types of delinquency, however, together with variations in the seriousness with which these behaviors are regarded by the law and wide variation in the size of correlations between behaviors, suggested that it might be possible to find clusters of related items which would prove to be theoretically interesting. A preliminary cluster analysis confirmed that this was the case and suggested further that factor analysis was a promising model for this purpose.[5]

The correlation matrix was factor analyzed, using the principal axis method and entering the highest column correlations in the main diagonal. Following a modification of Wrigley's criterion for when to stop factoring, five factors were extracted and rotated using Varimax.[6] The first unrotated factor accounted for much more of the common variance than did any succeeding factor, again pointing to the existence of a general delinquency trait, but there was ample evidence from the factor analysis that somewhat specialized adaptations also existed within our population.

Communalities of the 37 behavior items for the five-factor solution and their loadings on each of the factors are presented in Table 1.

Factor I is essentially a conflict factor, its highest loadings being individual fighting (.79), group fighting (.76), carrying concealed weapons (.67), and assault (.67). None of these behaviors loads highly on any other factor. It will be noted that robbery, theft, public nuisance, and statutory rape also load fairly highly on this factor, ranging from .51 to .40, but these items have similarly high loadings on other factors. Factor I may be more precisely characterized as consisting of conflict offenses, both acquisitive and destructive. It is most incompatible with work experience, team sports, attempted suicide, and narcotics involvement, as indicated by negative loadings of these items.

Factor II has highest loadings of individual sports (.71), team sports (.68), social activities (.60), and gambling (.48). Other loadings higher than .40 are obtained for joy riding, truancy, and hanging. The latter behaviors have similarly high loadings on other factors, however. This factor may be characterized as a configuration of stable corner activities. No seriously delinquent behavior loads highly on it, though several minor types of delinquency also load moderately on this factor.

Factor III is difficult to characterize. It has as its highest loading items sexual intercourse (.77), statutory rape (.68), petting (.67), signifying (.53), hanging on the corner (.44), and the use, buying, and selling of alcohol (.39). It also has the only moderately high loading of any factor for work experience (.36), accounting for nearly all of this item's common variance. We have chosen to call this a stable sex pattern and to regard the loading of work experience as a further indication of a type of relatively adaptive behavior which is represented by the factor.

Factor IV is characterized by high loadings for quite different sex behaviors, namely homosexuality (.53), having an illegitimate child (.50), and common law marriage (.48). Additionally, it has as its highest loading the use, buying, and selling of narcotics (.56) and of marihuana (.55). And, this factor accounts for virtually all of the low communalities of suicide and pimping, with loadings of .36 and .27, respectively. In striking contrast with Factor III, work experience has a loading of only .06, sex intercourse .17, statutory rape .18, petting .01, hanging .00, and signifying .00 on Factor IV. The combination of narcotics involvement, the "deviant" sex behaviors, and attempted suicide leads us to identify this factor as retreatist.

Factor V includes the highest loadings of auto theft (.69), driving without a license (.65), public nuisance (.58), theft (.53), use, buying, and selling of alcohol (.48), and running away from home (.44). In addition, this factor includes moderately high loadings for joy riding (.41) and truancy (.39). The versatility of this combination of offenses within a variety of institutional contexts (e.g., the "institutions" of property, school, and family, and in general the maintenance of public order) leads us to characterize this factor as an auto-theft-authority-protest pattern. As an abbreviated description we shall refer to Factor V as representing authority protest.[7]

Because we wished to maximize the independence of factors for purposes of etiological inquiry, it was decided to eliminate from factor scoring all items which were not relatively "pure" on a given factor. Factor loadings of .40 or higher or nearly equal smaller but highest loadings for an item on more than one factor were considered sufficient to eliminate an item from a given factor score. Scoring involved summing of the standard score, weighted by the factor loading of each item for each boy.

Behavior Factor Profiles of 16 Gangs

We turn now to an examination of the mean factor scores of our 16 gangs, presented in Table 2.

Because of the method of sample selection it is clear that we must be extremely cautious in "hypothesizing" relationships between variables within our population on the basis of competing theories of delinquent subcultures. Discussion in this paper will be limited to the relation of race to group mean factor scores. Following Cloward and Ohlin, it was our expectation that Negro gangs would have higher conflict factor scores than would our white gangs. That is, both legitimate and illegitimate economic opportunities seem objectively more limited for Negroes than for whites--hence, the expectation of greater conflict orientation among Negro than among white gangs.[8]

Findings in Table 2 may be summarized as follows: (1) The most conflict-oriented gangs are Negro. All six gangs with mean scores above the total population mean are Negro. The mean rank of Negro gangs on Factor I is 7.7 (out of 16) as compared with 10.2 for white gangs. There is much variation in conflict orientation among Negro gangs, however. Three of the four least-conflict oriented gangs also are Negro.

(2) The range of between-group variation within this population is greatest for Factor I (conflict) and least for Factor V (authority protest).

(3) The largest difference between contiguous groups in rankings occurs between the highest and next highest ranking groups on Factor IV (retreatism). This difference is 1.53 standard units, while the difference between the second highest group on retreatism and the lowest is only 1.12 standard units. The magnitude of these differences makes comparison of mean ranks of Negro and white groups relatively meaningless for this factor.

(4) Negro gangs are, on the average, also higher on Factors II and III, with mean ranks of 7.5 and 7.8, respectively, compared with mean ranks of 10.8 and 10.0 for white gangs.

(5) Only on Factor V do white gangs clearly rank higher than Negro gangs. Four of the five white gangs have positive group means for this factor and three of the four highest ranking groups are white. Mean rank for white gangs is 5.6, compared to 9.8 for Negro gangs.

(6) With one exception, individual group profiles are positively correlated with one another, but vary greatly and inconsistently in the relative elevation of different factors. When rank order correlations between the ranks of the 16 gangs on each of the five factors are computed, only Factor II rankings fail to correlate consistently in the positive direction with the rankings of other factors. No positive correlation with Factor II rankings is high, and its highest correlation is negative, with Factor V. Groups high on conflict tend also to be high on Factors III, IV, and V. Despite generally positive correlations, however, much variation in the relative elevation of any group on any given factor remains unexplained by its elevation on any other factor.

Discussion

Data presented in this paper are only indirectly relevant to the issue of the existence or the nature of delinquent subcultures. We have, after all, investigated the behavior of individuals rather than groups. In order to assess subcultural theory on the basis of these data, we must assume the relevance of the groups to which these individuals belong for understanding their behavior. Such an assumption seems warranted in view of the methods of selection of boys for study and the group context within which the raters (detached workers) know the boys. Nevertheless, caution is necessary until other data are interpreted, as will become clear in the following comments.

The factors are interpretable in terms of patterns of subcultural delinquency discussed in the literature. Conflict and retreatism emerge as fairly distinct emphases in terms of factor structure, but criminal behavior does not. Clearly, however, neither conflict nor retreatism are exclusive emphases of the gangs studied. They are positively correlated with each other, both when scores of individuals are correlated and when the rank order of group mean scores are correlated.[9]

We note that no clearly "criminal subculture" factor was extracted by our procedures. Although the five factors are not "equally criminal," as an examination of Table 1 reveals, the variance of criminal behaviors is spread over all of the factors to a greater extent than either conflict or retreatist behaviors.

Criminal behavior appears to be related to each of the other "delinquent" patterns in ways which may be consistent with the dominant emphases in these patterns, though our evidence is not complete enough to document completely such a conclusion. We know from observational data, however, that strong-arming and robbery (both components of the robbery item) cliques sometimes emerge within gangs high in conflict orientation. Similarly, the tough conflict-oriented boys sometimes display and utilize their neighborhood

"rep" by charging small amounts from younger boys for "protection" or by "shaking down" paper boys. Criminal behaviors loading high on Factor IV, strong-arm robbery, pimping, and petty thefts, seem most appropriately characterized as "hustles" toward the acquisition of money to finance drugs, alcohol, and other kicks.[10]

The evidence presented here appears to argue for the existence of types of behavior which are common to all gangs. It has been suggested that these items may constitute a "parent delinquent subculture," out of which the more specialized delinquent adaptations emerge. Neither the validity nor the utility of such a concept can be assessed on the basis of data presented in this paper, but the data appear to be consistent with such a formulation. The extent to which specific delinquency emphases come to characterize a gang at any point in time, and if they do, just how such specialization comes about is unclear from these data, and must remain so until further analyses can be undertaken. Hopefully, further study of such variables and processes as the reaction of these lower class gang boys to a variety of institutional contexts (involving, for example, relations in the family and at school, and the values of private property and keeping the peace); the study of interpersonal relations within the gang as well as between gangs, and relations between the gang and the external world; and the study of other variables specified in competing theories attempting to account for subcultural delinquency will provide answers to some of these questions.

The full implications of our findings for subcultural theories of juvenile delinquency cannot be assessed until the data are viewed in combination with observational data from detached workers and our own staff, and in terms of analyses of their relation to etiological variables specified by the theories. The former add the richness and detail of situational and group-process determinants which the ratings employed in the factor analysis miss. At the same time, without the more systematic and "objective" ratings, one cannot be certain as to the representativeness of his observations or his objectivity in recall and choice of behavior reported. Further analyses of the data are presently underway in our research program.

Footnotes

*The research reported in this paper is made possible by grants from the National Institute of Mental Health (M3301) and the Ford Foundation. The authors wish to express their gratitude to Robert A. Gordon for aid in the statistical analysis and for critical review of earlier versions of the paper. Our thanks also to the Program for Detached Workers of the YMCA of Metropolitan Chicago, without whose cooperation the entire research enterprise would be impossible.

[1]Albert K. Cohen, Delinquent Boys: The Culture of the Gang (Glencoe: The Free Press, 1955); Albert K. Cohen and James F. Short, Jr., "Research in Delinquent Subcultures," The Journal of Social Issues, XIV, 3 (1958), pp. 20-37; Richard A. Cloward and Lloyd E. Ohlin, Delinquency and Opportunity: A Theory of Delinquent Gangs (Glencoe: The Free Press, 1960); Walter B. Miller, "Lower Class Culture as a Generating Milieu of Gang Delinquency," The Journal of Social Issues, XIV, 3 (1958), pp. 5-20; H. H. Bloch and Arthur Neiderhoffer, The Gang: A Study in Adolescent Behavior (New York: Philosophical Library, 1958).

[2]See Cohen and Short, ibid., and James F. Short, Jr., "The Sociocultural Context of Delinquency," Crime and Delinquency (October, 1960), pp. 365-375.

[3]See, for example, Edward E. Schwartz, "A Community Experiment in the Measurement of Juvenile Delinquency," Yearbook of the National Probation Association, 1945, pp. 157-82; Alfred J. Kahn, Police and Children, New York Citizens' Committee on Children of New York City, 1951; Austin L. Porterfield, Youth in Trouble (Fort Worth: Leo Potisham Foundation, 1946); James F. Short, Jr., and F. Ivan Nye, "Reported Behavior as a Criterion of Deviant Behavior," Social Problems (Winter, 1957), 5, 207-213.

[4]See Ray A. Tennyson, "Detached Workers as Sources of Data," paper read at the annual meetings of the Society for the Study of Social Problems, August, 1960 (dittoed).

[5]See the writings of Louis L. McQuitty, e.g. "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies," Educational and Psychological Measurement, 17, 2 (Summer 1957), pp. 207-229; "Hierarchical Syndrome Analysis," Educational and Psychological Measurement, 20, 2 (Summer, 1960), pp. 293-304; "A Method for Selecting Patterns to Differentiate Categories of People," Educational and Psychological Measurement, 21, 1 (Spring 1961), pp. 85-94.

[6]Kenneth I. Howard and Robert A. Gordon, "An Empirical Note on the Number of Factors Problem in Factor Analysis," submitted for publication, July, 1962.

[7]We note the similarity of authority protest to the delinquent subculture described by Cohen. Later study will be concerned with the extent to which such activities represent participation in a subculture which is more amorphous and less specialized than the varieties of delinquent subcultures discussed in the literature. Such a generalized subculture has been hypothesized by Cohen and Short as a "parent delinquent subculture" from which the more specialized varieties develop. Cohen and Short, op. cit.

[8]See Cloward and Ohlin, op. cit.

[9]When factor scores are intercorrelated for individual boys, the resulting matrix also is positive, with product-moment correlations ranging from .06 to .47. Differences appear also between Negroes and whites in the intercorrelation of factor scores. These findings will be developed further in another paper.

[10]Cf. Harold Finestone, "Cats, Kicks. and Color," Social Problems, Vol. V, No. 1 (July, 1957), pp. 3-13.

Table 1

Loadings on All Factors and Communalities--
Five Factor Solution

Item	Communalities	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Group fighting	73	76	13	-30	-06	-19
Individual fighting	72	79	10	-22	-09	-19
Sex intercourse	70	26	02	-77	-17	-08
Statutory rape	66	40	01	-68	-18	-09
Public nuisance	65	43	34	-05	-10	-58
Theft	62	44	32	03	-22	-53
Assault	61	67	17	-03	18	32
Alcohol	60	36	06	-39	-26	-49
Auto theft	58	26	07	-10	-14	-69
Concealed weapon	56	67	05	-22	-13	-21
Individual sports	56	01	71	-13	-17	-07
Petting	55	12	22	-67	-01	-18
Robbery	51	51	09	-04	-35	-33
Driving w/o license	50	13	11	-17	-14	-65
Social activities	48	22	60	-26	06	06
Team sports	48	-06	68	-11	-02	03
Pot	46	29	10	-11	-55	-23
Hanging	42	07	40	-44	03	-25
Narcotics	42	-01	-08	04	-56	-32
Signifying	42	21	31	-53	04	-05
Truancy	41	26	43	04	09	-39
Joy riding	39	09	45	-10	-01	-41
Homosexuality	37	05	-10	03	-53	-27
Gambling	35	19	48	-20	-10	-17
Gang bang	29	32	08	-08	-28	-32
Illegitimate child	29	14	06	-15	-50	04
Runaway	28	27	10	02	-06	-44
Domestic chores	27	31	36	-17	11	02
Common law marriage	25	06	06	-07	-48	05
Work experience	18	-09	17	-36	-06	-10
Forgery	16	24	03	-09	-06	-30
Attempted suicide	14	-02	00	03	-36	-06
Bribery	11	03	-02	-10	-10	-29
Pimping	11	14	02	-07	-27	-11
Rape	10	18	-16	-05	-04	-19
Arson	04	00	-11	-01	-04	-15
Homicide	03	12	-03	-01	-11	-03

Table 2

Mean Behavior Factor Scores for 11 Negro and 5 White Gangs, Ranked by Size of Group Means

Behavior	Group Identification Number and Race															
Factor 1 (Conflict)	03 N 2.24	15 N 1.97	20 N .41	05 N .33	09 N .25	02 N .01	24 W -.05	18 W -.43	13 W -.44	23 N -.51	01 N -.68	22 W -.78	21 N -1.02	10 N -1.44	17 W -1.71	11 N -1.94
Factor 2 (Stable Corner Boy Activity)	23 N 1.43	20 N .90	09 N .85	21 N .75	13 W .51	15 N .39	17 W .23	10 N .09	03 N .08	02 N -.34	05 N -.48	22 W -.49	11 N -.62	18 W -.73	01 N -.78	24 W -.89
Factor 3 (Stable Sex-Maturity)	09 N .85	15 N .82	03 N .71	05 N .54	10 N .52	13 W .02	24 W .01	20 N -.03	02 N -.17	22 W -.24	18 W -.26	11 N -.31	01 N -.82	23 N -1.14	21 N -1.15	17 W -1.30
Factor 4 (Retreatist)	24 W 2.16	09 N .63	20 N .51	02 N .10	05 N .00	22 W -.04	23 N -.09	15 N -.10	03 N -.13	18 W -.23	01 N -.27	21 N -.31	13 W -.41	10 N -.45	11 N -.48	17 W -.49
Factor 5 (Parent Delinquent Subculture)	24 W .80	15 N .75	18 W .42	13 W .39	11 N .30	20 N .26	03 N .24	22 W .12	02 N -.11	09 N -.13	23 N -.18	17 W -.21	05 N -.34	01 N -.49	21 N -.49	10 N -.65

THEORY CONSTRUCTION AND PREDICTION IN JUVENILE DELINQUENCY*

Leroy C. Gould and Clarence Schrag

University of Washington

The Problem

The maturity of an empirical discipline is often evaluated in terms of its predictive capacity. If predictions are consistent with the results of repeated observation and experiment, this is powerful evidence in support of the generalizations or theories employed. In its early stages of development, a discipline makes predictions from generalizations concerning varied and discrete items of information. Integration of such generalizations under comprehensive theories occurs later.

Viewed from this perspective criminology is a young science. It has made considerable progress in establishing methods for selecting and combining predictive variables, comparing statistical and clinical predictions, revising predictive instruments on the basis of experience in their use, and assessing the impact of alternative treatments applied to a given population of offenders. But prediction and theory construction are still regarded as separate operations. Prediction consequently has little influence on criminological theory, and theory is infrequently employed in the development of predictive devices.

The problems of theory construction and prediction are nevertheless similar. We begin in both cases with a set of variables that are presumably related to delinquency. The task is then to determine the interrelations among the variables and their joint effect on delinquency. In prediction, an equation asserts the estimated probability of delinquency by combining variables or assigning weights based on previously observed relationships. Theory construction involves the formulation of a minimum set of postulates (law-like assumptions concerning relations among a few key variables) from which the remaining interrelations can be logically derived. While prediction can often be achieved in the absence of theory, a valid theory will provide the information needed for prediction. Thus, the requirements for prediction and theory are comparable but more rigorous in the case of theory.

A closer rapprochement between theoretical and predictive inquiries would be beneficial to both. Prediction affords the supreme test of a theory's pragmatic effectiveness. Theory, on the other hand, systematizes the generalizations employed in prediction and elucidates the deductive implications of the system. The two procedures should therefore complement each other.

How to attain a more effective alliance between theory and prediction is the topic of this paper. It compares delinquent and nondelinquent boys, develops an elementary theory of delinquency, and illustrates the use of the theory in delinquency prediction. Although the illustrations are simple and quite restricted, the same analytical techniques are applicable to more complex and comprehensive theories and predictive devices.

The Data

Data were obtained on 195 boys attending two high schools located in the area of highest delinquency within a large Western city. The area, like most of those having excessive amounts of delinquency, is characterized by unemployment, unskilled workers, broken homes, ethnic minorities, school retardation, rental residences, buildings in need of repair, multifamily dwellings, etc. However, the two schools have a heterogeneous student body, including representatives of every social class and ethnic minority found in the city. The group studied is comprised of all delinquent boys and a random sample of the non-delinquent boys attending the two schools (117 delinquents, 78 non-delinquents, and a total population of 1700 boys.) School and court files were available. In addition, the subjects completed questionnaires. Variables investigated include age, race, social class, mobility, conduct, educational and occupational aspirations, normative attachments, perception of social opportunities, delinquent contacts, motivation for achievement, punitiveness, and self-conceptions. These are some of the variables identified in the literature as being importantly associated with delinquency.

* This paper is based on a larger study financed by the Ford Foundation and reported in Leroy C. Gould, Delinquency and Community Opportunity Structure, unpublished M.A. Thesis, University of Washington, 1961, and Delbert S. Elliott, Delinquency, Opportunity, and Patterns of Orientations, unpublished Ph.D. Thesis, University of Washington, 1961.

Clarence Schrag is primarily responsible for the axiomatic formulation of the theory reported here and for the resulting predictive instrument.

This paper reports a few of the findings that can be integrated within a theoretical system having predictive utility. The theory was constructed independently of the data and was suggested by studies of youth culture. These studies reveal similar results concerning the endorsement of middle class goals among delinquents and non-delinquents, their perception of opportunities for achieving these goals, their motivation for achievement, and the frequency of their contacts with delinquents and adult offenders. The results of the present study indicate that these variables are interrelated in a logically consistent manner and that most of their interrelations can be derived from a single set of postulates.

Matrix of Interrelations: A Theoretical Model

An empirical theory has three distinguishing features. First is a vocabulary listing the relevant variables or concepts and indicating the theory's scope and focus. Second is a dictionary connecting the variables or concepts with the data of observation and experience. Connections between theoretical terms and observed data are established by means of operational definitions and rules of correspondence. Third is a grammar comprised of rules governing the formation of sentences or assertions within the theory. The grammar specifies how the terms of the theory may legitimately be used and communicates the logical processes by which theorems may be derived from the postulates. If a theory is to be used for pragmatic purposes, it is necessary that these three components be clearly enunciated.

Vocabulary: The vocabulary of the present theory includes the concepts delinquency, delinquent contact, endorsement of middle class goals, perceived opportunity, and achievement motivation. The theory applies to male urban youth of high school age and excludes deviant behavior that results from accident or mental illness.

Dictionary: Each of the concepts of the theory is treated as a dichotomous variable. Operational definitions are applied as follows:

1). A subject is delinquent if he has a juvenile court record. If there is no record, he is nondelinquent.

2). He has high delinquent contact if he admits personal acquaintance with ten or more delinquents. If acquainted with less than ten delinquents, he has low delinquent contact.

3). He has high goal endorsement if he reports that going to college is important for one's success in life. Otherwise he has low goal endorsement.

4). He has high perceived opportunity if he states that he is likely to go to college. Otherwise he has low perceived opportunity.

5). His rating on achievement motivation is high or low depending on his response to Aronson's graphic test of need for achievement. In the test, figures are flashed on a screen in view of the subject who is later asked to reproduce them. Scoring is based on the manner in which the figures are reproduced, the amount of detail, etc.

Several other procedures for observing the concepts were also employed. More elaborate definitions therefore could be utilized in the theory. They might produce a more comprehensive and powerful theory. In the present analysis, however, the above definitions are retained because their simplicity makes it easy to compare the logical and the empirical structure of the theory. What is the best definition, in any case, depends upon the kind of analysis to be made of the data.

Grammar: The theory's grammar is communicated by its postulates. Four postulates are needed to connect the five concepts. They were drawn from previous studies and in most cases assert relationships that have been frequently observed:

I. Goal endorsement has a positive relationship with perceived opportunity.

II. Perceived opportunity has an inverse relationship with delinquent contact.

III. Delinquent contact has a positive relationship with achievement motivation.

IV. Achievement motivation has a positive relationship with delinquent behavior.

Each of the postulates is a two-valued statement claiming that a certain kind of empirical relation holds true for the two concepts involved. Since the theory contains five concepts, there are ten combinations of these concepts taken two at a time. To express these relations ten two-valued statements are required. Thus, six theorems must be deducable from the postulates if the theory is logically valid.

Before the theorems can be derived, the meaning of the relationships asserted by the postulates must be stipulated in greater detail. Various kinds of relations are permissible. In the absence of previous tests of the theory, decisions regarding the nature of the relations are perhaps a matter of intuitive judgment. Since the postulates were drawn largely from studies using correlation techniques, however, the decision was to consider the relations as being symmetrical and transitive. A symmetrical relation is one in which the statement that A is related with B implies that B is also related with A. Transi-

tive relations are those in which the statements that A is related with B and that B is related with C implies that A is also related with C.

These assumptions enable us to state the following theorems:

V. Goal endorsement has an inverse relationship with delinquent contact (Derived from posulates I and II).

VI. Perceived opportunity has an inverse relationship with achievement motivation (Postulates II and III).

VII. Delinquent contact has a positive relationship with delinquent behavior (Postulates III and IV).

VIII. Goal endorsement has an inverse relationship with achievement motivation (Postulate I and Theorem VI).

IX. Perceived opportunity has an inverse relationship with delinquent behavior (Postulate II and Theorem VII).

X. Goal endorsement has an inverse relationship with delinquent behavior (Postulate I and Theorem IX).

Regardless of the order or method in which the theorems are derived, their claims are consistent. Together with the postulates, they cover all possible combinations of the concepts. This demonstrates the logical validity of the theory. There remains, of course, the question of its empirical adequacy.

Matrix of Interrelations: The Empirical Evidence

The distribution of proportions for all combinations of variables, estimated for the total population by extrapolation of the observed data, is shown in Table 1. Arrows indicate the relationship claimed by the postulates and theorems. Note that in every instance the theoretical claim holds true empirically, although in some cases the measure of association is relatively low.

The table shows also that delinquency is not equally associated with each of the other variables. For the total population, the delinquency rate is .07. By contrast, the rate is .29 for subjects having high delinquent contact and .04 for those having low contact. Again, subjects with high perceived opportunity have a rate of .03, whereas those with low opportunity have a rate of .13. High and low achievement motivation and high and low goal endorsement reveal rates of .09, .05, .06, and .10, respectively. These findings suggest that in delinquency prediction the major contribution will be made by contact and opportunity.

Furthermore, the findings indicate that certain combinations of variables are infrequently observed. Low goal

Table 1

MATRIX OF INTERRELATIONS: EMPIRICAL DATA *
Each entry lists estimated proportions of total population.

	Perceived Opportunity		Delinquent Contact		Achievement Motivation		Delinquent Behavior	
	High	Low	High	Low	High	Low	Del.	Non-Del.
Goal Endorsement: High	.55	.26	.07	.74	.35	.46	.05	.76
Goal Endorsement: Low	.05	.14	.06	.12	.08	.11	.02	.17
Perceived Opportunity: High			.05	.54	.25	.35	.02	.59
Perceived Opportunity: Low			.08	.32	.10	.22	.05	.35
Delinquent Contact: High					.08	.05	.04	.10
Delinquent Contact: Low					.35	.52	.03	.84
Achievement Motivation: High							.04	.39
Achievement Motivation: Low							.03	.54

* Arrows indicate relationships implied by the theory. Single arrows specify theorems; double arrows specify postulates.

endorsement, for example, occurs rarely in combination with high perceived opportunity. Likewise, high delinquent contact is rarely combined with high perceived opportunity, and it occurs infrequently in combination with low achievement motivation. Apparently social processes in the community or in the subjects themselves operate so as to inhibit these particular combinations. What the processes are and how they operate is a subject for further investigation.

The general finding is that the population studied is distributed among the combined variables in accordance with the claims of the theory. To use the theory for prediction, however, requires that we know the relative distribution of delinquents and non-delinquents among the different combinations. If the ratio is the same in the various combinations, the theory has little utility in predicting delinquency. The implication is that the ratio will fluctuate when different variable are combined and that the fluctuation will be consistent with the claims of the theory. Table 2 provides the distribution of delinquents.

		Contact: High	Contact: Low
Opportunity	High	High contact-High opportunity (Group A)	Low contact-High opportunity (Group B)
	Low	High contact Low opportunity (Group C)	Low contact-Low opportunity (Group D)

According to the theory, the ratio of delinquents to nondelinquents should be highest in group C and lowest in group B. The observed ratios are 1 to 1.6 in group C, 1 to 6.8 in group A, 1 to 13 in group D, and 1 to 57.1 in group B. The difference between groups B and C is in accord with the theory. Although the theory makes no explicit claim regarding groups A and D, the difference in ratios is consistent with the finding that delinquent contact has a higher association with delinquency than does perceived opportunity.

Among the variables one interacts with others in certain cases so as to contradict the theory. This variable is goal endorsement. It has an inverse relationship with delinquency. But it shows a positive relationship with delinquency when combined with high delinquent contact (when combined with low

Table 2

MATRIX OF INTERRELATIONS: EMPIRICAL DATA FOR DELINQUENT POPULATION
Each entry lists observed proportion of delinquent population.

		Perceived Opportunity		Delinquent Contact		Achievement Motivation	
		High	Low	High	Low	High	Low
Goal Endorsement	High	.24	.48	.32	.39	.42	.30
	Low	.00	.28	.21	.08	.14	.15
Perceived Opportunity	High			.10	.14	.16	.08
	Low			.43	.33	.39	.37
Delinquent Contact	High					.30	.23
	Low					.26	.21

When two variable are combined, the effect is to modify the ratio of delinquents to nondelinquents. To illustrate, let us examine the combination of delinquent contact and perceived opportunity. Four groupings are produced as follows:

delinquent contact, it operates according to the theory). This means that a special rule will have to be adopted when goal endorsement is used for predictive purposes.

Test of Predictions Based on the Theory

When the four dichotomous variables theoretically related to delinquency are combined simultaneously (four at a time) there are sixteen possible combinations. Prediction involves, first, estimating the relative delinquency rates in the different combinations and, second, arranging the combinations in a sequential order that is consistent with their ascending or descending delinquency rates. Since the postulates and theorems deal with only two variables at a time, we need another rule to establish priorities among the combinations of four variables. Various kinds of rules have been suggested for this purpose. One suggestion is to use a multiple regression technique in estimating the differential contributions of the variables. Another is simply to assume that delinquency rates are negatively correlated with the proportions of the total population that are found in the different combinations. The rule employed in the present analysis, however, is based on the marginal frequencies of the original matrix indicating the relationship between delinquency and each of the other variables. The rule asserts that in any combination the greatest contribution to delinquency is made by contact, the next greatest by opportunity, the next by achievement motivation, and the least by goal endorsement. This rule establishes a system of priorities among the variables so that their combinations can be arranged in a predetermined sequence.

The procedure for employing the rule is illustrated below. First the combination theoretically having the highest delinquency rate is recorded (High Contact, Low Opportunity, High Achievement, High Goals). There are four combinations that deviate from this pattern on only one variable. These are listed next. Then the six combinations deviating on two variables are listed, followed by the four that deviate on three variables, and finally the one combination that deviates on all four variables. The result is a complete listing of combinations ordered roughly in terms of their descending delinquency rates.

The priority rule provides for further ordering of the combinations within each of the sets listed. Among the combinations that deviate from the delinquency pattern on only one variable, for instance, the one deviating on the weakest variable (Goals) is listed first, then the one deviating on the next weakest (Achievement), and so on. Table 3 presents the observed delinquency rates for the combinations ordered in this manner. It is clear that the rule provides for an appropriate arrangement of the empirical findings.

Some criminologists have argued that separate theories of delinquency will have to be constructed for different ethnic groups and social classes. The implication is that these groups respond differently to social situations and that different causal processes are in effect. If this is the case, then our predictions, although fairly accurate for the total population, may be inadequate for certain populations segments. For this reason separate prediction tables were constructed for the different ethnic and social groups in the study population. No appreciable differences were noted in the accuracy of prediction. The rate of

Table 3

TEST OF PREDICTIONS ON TOTAL POPULATION

Contact	Opportunity	Achievement Motivation	Goal Endorsement	Delinquents	Non-dels.	Projected Total	Prop. Dels.
High	Low	High	High	13	0	13	1.00
High	Low	High	Low	12	1	32	.38
High	Low	Low	High	13	1	33	.39
High	Low	Low	Low	12	2	53	.23
High	High	High	High	10	3	71	.14
High	High	High	Low	0	1	20	.00
Low	Low	High	---	21	12	265	.08
High	High	Low	High	2	0	2	1.00
High	High	Low	Low	0	0	0	--
Low	Low	Low	---	18	13	282	.06
Low	High	High	---	9	16	334	.03
Low	High	Low	---	7	29	596	.01

delinquency, of course, is much higher for Negroes than for Caucasians and for lower class boys than for middle class boys. But the prediction data suggest that these variations can be explained largely in terms of the relative concentrations of these groups within the different combinations of variables listed in the theory. Table 4 gives the relevant data for ethnic minorities, excluding orientals, in the study population.

What about the remaining variables included in the study? It appears that some of them can be added to the theory, thereby increasing its comprehensiveness and possibly its predictive efficiency. Delinquents, for example, tend to rate security as being more important to them than happiness while nondelinquents attach greater importance to happiness. Again, delinquents often perceive social norms as idealistic prescriptions that nearly everyone violates, whereas nondelinquents see the norms as actually influencing people's conduct. Important differences also are found in self-conceptions, punitiveness, and several other factors. Perhaps these can be incorporated within the theoretical framework. Some other variables probably can best be used as indicators of conditions under which the theory holds true. The specification of such conditions is of crucial significance in any theory that is to be used for pragmatic purposes. Thus far, criminology has paid little attention to the conditions under which theories operate.

Conclusion

Our objective in this report is to demonstrate that it is both possible and profitable to coordinate theory construction and prediction in empirical research. The elementary theory developed here is used for illustrative purposes and is by no means a finished product. So far as predictive efficiency is concerned, however, this theory compares favorably with most of the conventional devices employed in delinquency prediction.

Unless prediction is integrated with theory construction, there is little basis for evaluating alternative definitions of concepts or alternative formulations of postulates. Theory cannot be accepted on the basis of its logical validity alone. At the same time, predictions from independent and perhaps inconsistent generalizations are not likely to be effective when applied to different populations or to different time series. There is a sound basis for the assumption that theory will help to increase the reliability and validity of predictions. For these reasons we believe that criminological research would produce more cumulative and verifiable knowledge if theoretical and predictive frameworks were developed jointly.

Table 4

TEST OF PREDICTIONS ON ETHNIC MINORITIES - EXCLUDING ORIENTALS

Contact	Opportunity	Achievement Motivation	Goal Endorsement	Delinquents	Non-Dels.	Projected Total	Prop. Dels.
High	Low	High	High	2	0	2	1.00
High	Low	High	Low	3	0	3	1.00
High	Low	Low	High	4	0	4	1.00
High	Low	Low	Low	6	1	26	.23
High	High	High	High	5	1	25	.20
High	High	High	Low	0	0	0	--
Low	Low	High	---	10	6	132	.08
High	High	Low	High	1	0	1	1.00
High	High	Low	Low	0	0	0	--
Low	Low	Low	---	7	5	109	.06
Low	High	High	---	1	2	42	.02
Low	High	Low	---	0	3	61	.00

IV

EVALUATION OF THE 1960 CENSUS

Chairman, Frederick F. Stephan, Princeton University

Page

The Accuracy of the 1960 Census Count - Joseph Steinberg, Margaret Gurney and Walter Perkins, U. S. Bureau of the Census..... 76

Measurement Errors in the 1960 Census of Population - Leon Pritzker and Robert Hanson, U. S. Bureau of the Census.... 80

Evaluation Studies of the 1959 Census of Agriculture - R. Hurley, T. Jabine and D. Larson, U. S. Bureau of the Census..... 91

Study of the Reliability of Coding of Census Returns - Herman H. Fasteau, J. Jack Ingram and Ruth H. Mills, U. S. Bureau of the Census..... 104

Discussion - Ansley J. Coale, Princeton University..... 116

THE ACCURACY OF THE 1960 CENSUS COUNT

Joseph Steinberg, Margaret Gurney
and Walter Perkins

I. Introduction

The single most important set of statistics obtained from the 1960 Decennial Census is the count of the total population of the United States and each of its subdivisions. One of the major objectives of a Census is to obtain a complete and unduplicated count of the population. Achieving this objective is difficult -- the high degree of population mobility, the length of the period of enumeration, the difficulty of locating some housing units, the multiple residences of some families, the problems of finding people at home in large cities and the problems of carrying out a large scale enumeration with a relatively inexperienced crew of enumerators, all cause a departure from the basic objective. With these many different sources of error possible, it is important to determine the degree to which we have achieved a complete and unduplicated count.

The purpose of this paper is to discuss some available information regarding the accuracy of the 1960 Census count. The data represent only some of the results that will be available and are considered here without the full and total examination which is called for in the final evaluation of the accuracy of the 1960 Census count. As separate components, these results will play a role in the final estimates. As yet they represent only part of the story which must be reviewed in order to determine the final estimates of the accuracy of the 1960 Census.

The test for completeness of enumeration of the 1950 Census rested both on the results of a Post Enumeration Survey and on the evidence of demographic analysis. [1/]

The Evaluation Program of the 1960 Censuses [2/] was designed to take account of what we learned from the 1950 PES. After discussion with a panel of expert consultants, a number of interrelated studies were developed to evaluate the accuracy of the count. It was decided that the studies should comprise:

1. Resurveys of samples of the population

2. A series of reverse record checks--that is, studies involving the determination of whether a sample of individuals selected from records could be found in the Census.

and 3. Analytical techniques--such as setting up an accounting of population growth, using Census counts and reported births, deaths and international migrations.

The final estimates will coordinate and reconcile results from all of these. At present, not all of the component results are available. For this and other reasons the present paper gives only the authors' personal views and are not official estimates of the accuracy of the 1960 Census count.

II. Measurement of Missed and Overcounted Units and Persons: Reenumerative Surveys

The reenumerative survey which dealt with under- and over-counts of housing units (and people in them) involved an area sample consisting of about 2,500 segments, with about 18,000 housing units. This sample was reinterviewed with an intensive procedure in the fall of 1960.

According to preliminary tabulations from this study, the 1960 Census missed about 1.9 million housing units--both occupied and vacant--this was about 3.2 percent. There were about 2.9 million missed persons in these missed units (1.6 percent). We estimate that about 400,000 to 600,000 housing units containing about 250,000 persons (including some units enumerated in the wrong ED) were enumerated in error in the Census. Thus, we estimate a net underenumeration of about 2.4 percent of the housing units, and within these housing units 1.5 percent of the enumerated population.

The estimates just given resulted from the following steps:

a. An intensive canvass of the area segments in 1959,

b. An allocation of units covered in the Census to these land areas, by a field visit in the summer of 1960, and finally,

c. An intensive canvass by specially trained enumerators in the fall of

---

1/ U.S. Bureau of the Census, Post Enumeration Survey; 1950. Bureau of the Census, Technical Paper No. 4, Washington, D.C. 1960

2/ Hansen, Morris H., Pritzker, Leon, and Steinberg, Joseph, "The Evaluation and Research Program of the 1960 Censuses" American Statistical Association, Proceedings Social Statistics Section, Washington, D. C., December 1959, pp. 172-180.

1960. In this intensive canvass, the enumerators attempted to identify all units that had been enumerated in each structure in the previous enumerations as well as to determine what the "true" situation was at the time of these enumerations. Thus, available for analysis of coverage errors in the Census are records that describe the housing units that were enumerated at the time of the Census, and the housing units that should have been enumerated.

Another study also had provision for finding missed units in the physical neighborhood of units which were in sample. Estimates from this study are lower than the figures cited, but the number of persons per occupied-unit missed is of the same order.

Let us consider a comparison of coverage errors for occupied units for the 1960 Census with the corresponding data from the 1950 PES (Post Enumeration Survey). For net underenumeration of occupied housing units, the rate in 1960 is 2.0 percent, as compared with 2.3 percent in 1950. For net persons missed in missed units, the rate is 1.5 percent in 1960, as compared with 1.3 percent in 1950.

These comparisons are made for occupied units only, since the estimates of the number of missed (or overenumerated) vacant units in both the 1950 and the 1960 surveys are subject to greater error. Greater errors for vacant units occur because:

1. It is difficult to determine whether a vacant structure meets the housing unit criteria, particularly when it is in poor repair.

2. Vacant housing units which are reached through other (occupied) units are easily missed.

3. It is virtually impossible to match vacant seasonal cottages, found in a sample segment, with listings in the Census.

The errors in the Census counts of housing units resulted from a number of causes: Among all missed housing units, vacant as well as occupied, about 60 percent or about 1.9 percent of the Census total, were in omitted structures. The remaining 40 percent (1.3 percent of the Census total) were missed units in enumerated structures.

Among the erroneously included housing units, about 13 percent were "nonexistent" units, resulting from misclassifications of units in the Census, fictitious enumerations, etc. The remaining 87 percent (less than one percent of the Census total) were divided into two classes:

1. Structures having all units erroneously included accounted for 52 percent of the overcounted units. This class includes single unit structures as well as multi-unit structures.

2. Multi-unit structures having one or more units correctly enumerated gave rise to the remaining 35 percent of the overcounted units.

Thus, for both under- and over-counted units, the preponderance of errors occurred when the whole structure was missed or counted in error.

III. Measurement of Missed and Overenumerated Persons in Enumerated Units

The net undercoverage in the 1950 Census was substantially underestimated by the Post Enumeration Survey taken in 1950. Further checking on the results indicated that little, if any, of this deficiency represented a failure in the PES to find living quarters that had been missed in the Census. Probably an important share of those left out of the Census consists of persons who have no regular place of residence or who have a very tenuous connection with what may be considered their residence. It is just this group that the 1950 PES also had difficulty in locating and counting.

Because of the deficiencies in the 1950 coverage estimates, the methods and procedures for evaluating coverage in the 1960 Census in enumerated units were strengthened in a number of respects. Probably the most important single improvement consisted in shortening the time gap between the evaluation and the Census. In 1950 the average interval between the original enumeration and the reenumeration was between four and five months. In 1960, this difference was more in the neighborhood of four or five weeks.

A second difference was the change from a dependent reenumeration in 1950 to a 1960 evaluation that was largely independent. A reenumerator in 1960 was given the address and the name of the head for each of the 15,000 sample units. This amount of information automatically told the reenumerators which units were considered vacant by the Census and which were considered occupied. Beyond this point, however, the reenumerator had no information as to who or how many persons had been counted by the 1960 enumerator. The independent interview produced many more differences as compared with the Census than would a dependent reinterview. Of course, when a reconciler returned at a later date to check the coverage differences some discrepancies turned out to be errors in the reenumeration rather than errors in the Census. It was the job of the reconciler to weed out the reenumeration errors so that the final estimates would reflect only coverage deficiencies in the Census.

A third respect in which the 1960 evaluation differed from the 1950 PES is that the effect of processing rules on coverage has been estimated in 1960. The evaluation of 1950 coverage stopped with what was recorded by the Census enumerator; no attempt was made to measure differences in coverage coming from punching and tabulating.

Evaluation procedures in 1960 have provided higher, and probably more reasonable, estimates of the numbers of missed persons in enumerated housing units than in 1950. Where the 1950 PES estimated less than a million missed persons in enumerated living quarters, the total corresponding estimate for 1960 is roughly two million persons; this, despite the fact that there is evidence that, at least in net effect, the 1960 Census did a better job in coverage than the 1950 Census.

The 1960 evaluation procedures showed that there was a field overcount of about 1.2 million persons in enumerated units. This leads to an estimated net undercount of about 800,000 in the field work. However, in the processing some rules were followed which had the effect of arbitrarily recording population for certain housing units that had been designated as occupied on the Census returns, but for which no data were readable to FOSDIC for individual household members. This was accomplished in each case by replicating information for another household. As a result of these rules, data for about 600,000 persons were replicated in the Census counts. Accordingly, we estimate a net undercount within enumerated units somewhere in the neighborhood of 200,000 persons, or about 0.1 percent. The 1950 PES on the other hand estimated a net field overcount of about 100,000 persons in enumerated living quarters.

IV. Reverse Record Checks

Having discussed some of the available information from our reenumerative surveys in the preceding paragraphs, let us turn to our reverse record checks. In brief, in order to establish an almost complete frame of the total population, we have drawn a sample from four basic sources:

1. The 1950 Census
2. A sample of the people found, according to the 1950 PES, to have been missed in the 1950 Census
3. Births occurring since the 1950 Census, and
4. Aliens registered on January 1960 as being residents of the United States.

From each of these sources a small sample was selected designed to develop information on the accuracy of the count for the particular group.

The final results are not yet available from these reverse record checks. However, we expect to find current addresses (or that sample people were deceased) for about 90 percent of the sample cases. No attempt was made prior to 1960 to test the degree of coverage of a Census by selecting a sample from other records. Therefore, these efforts represent a new approach and will be studied carefully in relation to the reenumerative surveys and demographic analysis.

V. Demographic Analysis

Trial calculations have been made by Donald Akers of the Bureau of the Census of the net Census undercount through the use of several analytic techniques. 3/ These trial calculations are labeled as experimental and not to be taken as official estimates of the Bureau of the Census. One of the analytic methods indicates that the absolute error of the count in the 1960 Census is about the same as that in the 1950 Census or that the net relative undercount dropped from 2.4 4/ to 2.1 percent. In analyzing the possible sources of error in each of the components used in the determination of the result, Akers indicates that the margin of error may be as much as 0.4 percent around the 2.1 percent.

In another method, using a battery of iterative techniques similar to those used by Ansley Coale in judging the 1950 Census 5/, Akers derived a result which estimates that the net relative undercount in the 1960 Census is about one-sixth less than in the 1950 Census.

VI. Summary, Conclusions, and Direction of Further Efforts

This paper reflects our personal opinions based on partial results. The results seem to suggest that through resurvey methods a net undercount of population of about 1.6 percent may be estimated, and through analytical methods between 1.7 and 2.5 percent. Considering the evidence available to us, we conclude that the most likely level of net undercount in 1960 is in the range of 1.6 to

---

3/ Akers, Donald S. "Estimating Net Census Undercount in 1960 Using Analytical Techniques", paper presented at the annual Population Association of America Meetings, May 5, 1962.

4/ "Minimum Reasonable" estimate, U.S. Bureau of the Census, Technical Paper No. 4, op cit, p. 6.

5/ Coale, Ansley J. "The Population of the United States in 1950. A Revision of Census Figures." J. Amer. Stat. Assn., Vol. 50, pp. 16-54.

2.0 percent of the population as compared to the "minimum reasonable" estimate in the 1950 Census of 2.4 percent. In absolute terms, this amounts to a net undercount in 1960 between 3 and 3½ million people.

For net underenumeration of occupied housing units our estimate is a rate of 2.0 percent in 1960 as compared with 2.3 percent in 1950:

Subject	Net Undercount			
	1960		1950	
	Percent	Number (millions)	Percent	Number (millions)
Population	1.6-2.0	3.0-3.5	2.4	3.7
Occupied housing units	2.0	1.1	2.3	1.1

We believe that the techniques used in the reenumerative surveys are close to a level of maximum intensity. It is possible that, through the development of more effective questions and more careful evaluation work in hotels and institutions, somewhat better results might be achieved in the next series of evaluations of a Decennial Census.

Acknowledgments

The Evaluation and Research Program of the 1960 Census required substantial contributions by a large number of people. The program is under the general direction of Conrad Taeuber, Morris H. Hansen, and William N. Hurwitz. Technical planning was the primary responsibility of Leon Pritzker and Joseph Steinberg, supported by other staff members. The operation of the program was the primary responsibility of Joseph Steinberg.

The authors wish to acknowledge especially the contributions to the studies described in this paper by Barbara Boyes, Ruth Mills, William Enright, and William Taylor.

****

MEASUREMENT ERRORS IN THE 1960 CENSUS OF POPULATION

Leon Pritzker and Robert Hanson
Bureau of the Census

(This paper presents the personal views of the authors, and not necessarily the official position of the Bureau of the Census.)

1. Mathematical model.

The following mathematical model, described in detail by Hansen, Hurwitz, and Bershad (2), has aided us in the design of our program of evaluation:

a. The basic postulates of the model — not attainable in the real world — are that the process of recording a response for any individual is 1) repeatable and 2) gives rise to a random variable whose value at trial t is not correlated with its value on any other trial (t + k) and whose expected value is constant over repeated trials.

b. Thus, we shall be dealing with a random variable, x_{jtG}, whose value is:

x_{jtG} = 1, if the recorded response classifies individual j in class x of some characteristic on trial t of a census conducted under a set of general conditions G. (x may denote the age-class "0 - 4," the income class "$10,000 and over," etc.)

= 0, if otherwise.

c. Then we can define the proportion classified in x in trial t of a census of n_t persons:

$$p_{tG} = \frac{1}{n_t} \sum_{j}^{n_t} x_{jtG} \qquad (1)$$

d. The task of evaluation, when viewed from this standpoint, is to obtain estimates of the mean square error of p_{tG} for the classes of the population for which census data are tabulated, for example, 5-year age classes, the categories into which income reports are combined, and the number of years of regular schooling completed. Thus:

$$MSE_{p_{tG}} = E(p_{tG} - \bar{U})^2 \qquad (2)$$

where the expectation is taken over trials and where $\bar{U}$ is the true proportion.

e. The mean square error can be divided into its two main components by subtracting and adding $E(p_{tG})$ inside the parentheses of equation (2). We define $E(p_{tG}) = P_G$ and:

$$MSE_{p_{tG}} = E(p_{tG} - P_G)^2 + (P_G - \bar{U})^2 \qquad (3)$$

$$= \sigma^2_{p_{tG}} + B^2_{p_{tG}} \qquad (4)$$

where the first term is the total variance of p_{tG} and the second term is the square of the bias of p_{tG}.

f. To express the variance of p_{tG} in a particularly fruitful way, let:

$$E(x_{jtG}) = P_{jG}, \quad 0 \le P_{jG} \le 1,$$

where the expectation is over trials for a fixed person.

Then, as a device to be used for separating the response variance from the total variance, let:

$$\hat{P}_{tG} = \frac{1}{n_t} \sum_{j}^{n_t} P_{jG} \qquad (5)$$

This is the average of the expected values for the sample drawn on trial t.

Then:

$$\sigma^2_{p_{tG}} = E(p_{tG} - \hat{P}_{tG} + \hat{P}_{tG} - P_G)^2$$

$$= E(p_{tG} - \hat{P}_{tG})^2 + E(\hat{P}_{tG} - P_G)^2 + 2E(p_{tG} - \hat{P}_{tG})(\hat{P}_{tG} - P_G) \qquad (6)$$

In equation (6), the first term is defined as the response variance, the second term as the sampling variance, and the third as the interaction.

g. Our interest in this paper lies, in addition to the bias, in the response variance. Let:

$$\sigma^2_{\bar{d}} = E(p_{tG} - \hat{P}_{tG})^2 \qquad (7)$$

$$d_{jtG} = (x_{jtG} - P_{jG}),$$

the deviation of the response recorded for individual j on trial t from the expected value of the responses over all trials.

Then the evaluation of the expected values indicated in equation (7) for a fixed sample size $n_t = n$ gives:

$$\sigma^2_{\bar{d}} = \frac{1}{n} E(d^2_{jtG}) + \frac{(n-1)}{n} E(d_{jtG}\, d_{ktG}) \quad (8)$$

h. The partition of the response variance provided by equation (8) gives a useful tool for the analysis of sources of unreliability.

Since $E(d_{jtG}) = 0$, $E(d^2_{jtG})$ is a variance, σ^2_{dG}. It is the basic trial-to-trial variability, averaged over individuals. By analogy with population sampling, it is defined as the "simple response variance." It can be shown that:

$$\sigma^2_{dG} = \frac{1}{N} \sum_j^N P_{jG}(1 - P_{jG}) \quad (9)$$

where N = the total number of persons in the population.

It can further be shown that:

$$\sigma^2_{dG} \leq P_G (1 - P_G) \quad (10)$$

And thus, we can define the "index of inconsistency":

$$I_{dG} = \frac{\sigma^2_{dG}}{P_G (1 - P_G)} \leq 1 \quad (11)$$

$P_G(1 - P_G)$ will be recognized as the "sampling variance" of an estimated proportion for a sample of one element in simple random sampling. In this model, however, $P_G(1 - P_G)$ also includes the "simple response variance."

We have found the index, I_{dG}, useful in determining the inherent reliability or "measurability" of any 0, 1 variate included in a census. Estimates of this index enable us to compare the inherent reliability of measurement of, for example, one five-year age class with any other five-year age class, or a broader age grouping with a narrower age grouping, or an age class with an income class.

Under certain circumstances, the more important component of the response variance is the second term of equation (8), reflecting the covariance between response deviations. Results bearing on this component are not yet available. However, a major effort is now in progress to provide estimates.

2. Discussion of two components of the mathematical model.

The two components of interest are:

$$B_{p_{tG}} = P_G - \bar{U} \quad (3a)$$

$$I_{dG} = \frac{\sigma^2_{dG}}{P_G (1 - P_G)} \quad (11)$$

In evaluating the national summary statistics of a census of population, of these two components the one of significance — except for very rare characteristics — is $B_{p_{tG}}$. The term, I_{dG}, is of interest in evaluating the precision of the census-taking process but not the accuracy of the statistics.

3. Estimators.

a. Bias.

Recall that

$$B_{p_{tG}} = P_G - \bar{U} \quad (3a)$$

For P_G, we write the census statistic itself, p_{tG}.

The problem, then, is to find a valid estimator for $\bar{U}$. This comes down to devising a method of measurement that can be agreed upon as "standard," or failing that, devising a "preferred" method that, by judgment or by test, is found to produce the required statistics with significantly smaller mean square errors than the statistics of a census. Then the estimator of $B_{p_{tG}}$ is:

$$b_{p_{tG}} = p_{tG} - p_{tG'} \quad (12)$$

where $p_{tG'}$ is an estimate produced by a standard or a preferred method.

The tables we have provided below give estimates of the bias, $b_{p_{tG}}$. These estimates come from the Evaluation and Research Program of the 1960 Census in which the following studies have been conducted to obtain estimates of the bias in national summary statistics:

(1) CPS - Census Match. (Study EP-23)

The Current Population Survey (CPS), conducted monthly by the Bureau of the Census, is regarded as a "preferred method," in relation to a census of population, for collecting data on the size and composition of the labor force. The data for a sample of about 8,000 house-

holds enumerated in both the April 1960 CPS and in the 1960 Census (conducted primarily in April 1960) have been brought together and analyzed. Some results of this match, as well as of a match conducted between the April 1950 CPS and the 1950 Census, are presented below.

(2) Population Content Evaluation Study. (Study EP-10)

This study was based on what we term an "intensive-interview" approach to obtaining census data. The "intensity" of an interview is a relative matter. We provide some notion of the "intensity" in Attachment A, in which are reproduced the questions employed in the inquiry to determine the age of each person in the sample for Study EP-10. We would assert that the interview designed for Study EP-10 was considerably more "intensive" than the interview designed for obtaining data for characteristics of the population in the Post-Enumeration Survey of the 1950 Census.

Data were obtained on the following characteristics by intensive interview: usual place of residence, age, place of residence on April 1, 1955 (providing a measure of mobility of the population), school enrollment and attainment, number of children ever borne by women who have ever been married, income from self-employment, and income from sources other than earnings. At the outset of each intensive interview, the interviewer also recorded the sex and color of the person for whom data was being obtained. In most cases the interviewer talked with the person himself when the person was an adult, and to the parent or guardian when the person was a child.

The results presented below are based on a sample of about 9,500 persons and reflect differences between the "best" answer obtained by the interviewers in Study EP-10, after reconciliation of differences between Census and EP-10 responses, and the Census responses recorded on the FOSDIC schedules of the 25 percent Census sample.

(3) Record checks and related studies.

The results are not yet available, but the Bureau of the Census expects to have, as part of the Evaluation and Research Program, estimates of bias in summary statistics from the following sources:

(a) Age and sex for children under 10 years of age from birth certificates.

(b) Occupation and industry of employed persons from employer records.

(c) Occupation, industry, and class of worker (the component of bias due to errors in coding) from a study being conducted by Fasteau, Ingram, and Mills (1).

b. Index of inconsistency.

Recall that:

$$I_{dG} = \frac{\sigma^2_{dG}}{P_G(1 - P_G)} \qquad (11)$$

We can write p_{tG} for P_G. Thus, the problem is to find an estimator for σ^2_{dG}.

Recall that:

$$\sigma^2_{dG} = E(d^2_{jtG}) \qquad (9a)$$

$$= E(x_{jtG} - P_{jG})^2$$

Now consider a theoretical repetition (t') of the Census (t) under the identical conditions, G. The repetition is theoretically independent. For any 0, 1 variate, for example age 15 - 19, the Census and its repetition would generate the following type of table:

Repetition \ Census	$x_{jtG} = 1$	$x_{jtG} = 0$	
$x_{jt'G} = 1$	a	b	a + b
$x_{jt'G} = 0$	c	d	c + d
	a + c	b + d	n_t

$$n_t = a + b + c + d = n_{t'} = n$$

Define the gross difference rate, g:

$$g = \frac{b + c}{n} \qquad (13)$$

It can be shown that under the specified conditions:

$$\sigma^2_{dG} = \tfrac{1}{2} E(g) \qquad (14)$$

Then, writing $\hat{I}_{dG}$ as an estimator of I_{dG},

$$\hat{I}_{dG} = \frac{g}{2p_{tG}(1 - p_{tG})} \qquad (15)$$

As pointed out by our colleague, Max Bershad, this term is equivalent to χ^2 for a two-by-two contingency table, as defined by Karl Pearson (3).

In practice the conditions do not hold and thus $\hat{I}_{dG}$ is a biased estimator of I_{dG}. In particular, there is a growing body of evidence that the division by 2 cannot be justified where the "repetition" is followed by reconciliation with the Census. In our judgment, the division by 2 has provided estimates that understate the response variability of the 1960 Census statistics. We believe, however, that $\hat{I}_{dG}$ is a useful estimator for helping us assess the relative consistency of recorded responses, as between characteristics and as between censuses. Biased estimates of I_{dG} will be available for selected characteristics, for the United States as a whole, from the following studies conducted as part of the Evaluation Program of the 1960 Census:

(1) CPS - Census Match. Study EP-23 — described above.

(2) Population Content Evaluation Study, Study EP-10 — described above.

(3) Replication Study. Study EP-18 This is a study based on the re-enumeration of 6,000 households included in the 25-percent sample of the 1960 Census. Results are not yet available.

4. Results and interpretations.

The results presented below are preliminary in three respects. First, there will probably be some corrections in the estimates themselves. Second, the estimates do not take account of the effect of errors in the coverage of the population and in the coverage of housing units on accuracy and reliability. Third, the estimates do not take account of the effect of nonresponse on accuracy and reliability.

The results are based on estimated "identical populations." For each characteristic studied, the "identical population" is projected from those persons (or housing units) for whom responses were recorded both in the Census and in the survey used as the standard to evaluate the Census.

a. Labor-force status. As indicated above, the standard for evaluating the Census is the CPS. Tables 1 and 2 present the results of two CPS-Census Matches. Table 1 permits us to compare the biases in the statistics for labor-force status of the 1960 Census with those of the 1950 Census. The statistics for females appear to have been uniformly improved in 1960 over 1950. The statistics for males also appear to have improved in 1960 over 1950 for unemployment. For employment in agriculture, the 1960 figures appear to be worse, and, for the other components, about the same. However, there are still estimates of bias in the 1960 Census statistics that some analysts would probably regard as important. For example, the male, civilian labor-force total in the 1960 Census is understated by more than 2 percentage points and the female, civilian labor force is understated by about 1 percentage point.

Table 2 provides a basis for comparing the data-collection procedures in the 1950 and 1960 Censuses by showing estimates of the indexes of inconsistency, $\hat{I}_{dG}$, for the labor-force status categories. The estimates are biased in that the CPS, at the time of each Census, has been taken to be the "repeated trial." We recognize that the CPS is not taken under the same conditions as the Census. In fact, we regard the CPS to be better for determining labor-force status. However, if the following two reasonable assumptions hold, we can conclude that the estimates of the index are underestimates:

(1) The covariances between the CPS and the Census response deviations are zero or positive.

(2) The simple response variances, σ^2_{dG}, of the CPS are less than the simple response variances of the Census for labor-force classifications.

There is some reason to believe that the 1960 CPS was of higher quality than the 1950 CPS. However, if we further assume that simple response variances of the CPS in 1950 are about equal to those of the CPS in 1960, we can make some rough inferences about the quality of the 1960 Census procedures as compared to the 1950 procedures. With possibly one or two relatively minor exceptions, the 1960 Census procedures appear to be at least the equal in reliability of the 1950 Census procedures. However, we can see by comparing Table 2 and Table 3 that the estimated indexes for labor-force characteristics are several times greater than for sex or color. Both the 1950 and the 1960 Census procedures for determining labor-force status generated a considerable amount of noise, i.e., response variance, particularly in the unemployment classification.

b. Other characteristics of the population. Table 3 presents estimates of the bias, $b_{P_{tG}}$, and of the index of inconsistency, $\hat{I}_{dG}$, for sex and for color. Table 4

presents the same types of estimates for age; Table 5, for mobility status; and Table 6, for educational attainment of the population 25 years old and over. Study EP-10 described above has been used as the source of both types of estimates for the 1960 Census. No corresponding estimates are available for sex and race in the 1950 Census. For age, mobility, and educational attainment, the Post-Enumeration Survey, or "PES" (4), has been used as the source of both types of estimates for the 1950 Census. All of the estimates — for 1960 and for 1950 — are undoubtedly biased. There are some a priori reasons for believing that Study EP-10 in 1960 was more "intensive" and had smaller mean square errors than the PES in 1950.

We offer the following comments about Tables 3 - 6:

(1) The reliability of the classification of the population by sex and by color was high in comparison to the reliability of classification of most other characteristics studied in censuses and surveys. However, there is some indication — to be investigated further — that there was a bias in the direction of understating the number of females and of nonwhites. (Recall that these estimates do not reflect coverage error or nonresponse.) The females are understated by 0.4 percent and the nonwhites by 1.7 percent, relative to the estimates of Study EP-10.

(2) The estimates of the index of inconsistency, $\hat{I}_{dG}$, indicate almost uniformly more reliable classifications in the 1960 Census than in the 1950 Census. As we have indicated earlier, the indexes are biased. It is within the realm of possibility that these results can be accounted for entirely by more reliable processes of classification in Study EP-10 than in the PES of the 1950 Census. It would be difficult for us, however, to believe that the improvement in our method of evaluation was so great as to mask a decline in the reliability of the classification processes in the 1960 Census itself. We would conclude that the quality of classification in 1960 was at least equal to or perhaps better than the quality in 1950 for age, mobility status, and educational attainment.

(3) Pending tabulation and analysis of the final results, we would like to withhold judgment concerning the estimates of bias, except for the following observations: First, the bias in the age class, 65 - 69, is consistent with the direction indicated by demographic analysis. In 1950, the PES result was not consistent. Second, the fact that the estimates of bias for mobility status in 1960 are generally higher than those in 1950 has to be viewed against the fact that the 1960 classifications cover a five-year period and the 1950, a one-year period. Third, the estimates of bias in educational attainment of the population 25 years old and over both for 1950 and for 1960 are generally consistent with the hypothesis of a net tendency to overstate educational attainment.

5. Concluding remarks.

The general method of measurement that we have illustrated has provided useful information to the Bureau of the Census. However, it has also been a source of great frustration. The index, I_{dG}, is a measure of the noise in the census-taking process. The bias, $B_{P_{tG}}$, is a measure of the misinformation in the process. The specific techniques developed thus far to estimate these quantities also have high noise levels and also give erroneous signals. The search for standards of measurement has a long way to go.

6. Acknowledgments.

The Evaluation and Research Program of the 1960 Census required substantial contributions by a large number of people. The program is under the general direction of Conrad Taeuber, Morris H. Hansen, and William N. Hurwitz. Technical planning was the primary responsibility of Leon Pritzker and Joseph Steinberg, supported by other staff members. The operation of the program was the primary responsibility of Joseph Steinberg.

The authors wish to acknowledge especially the contributions to the studies described in this paper by Donald E. Ball, Stanley Greene, Eli S. Marks, Ruth H. Mills, and Martha Thomas.

REFERENCES

1. Fasteau, H. H., Ingram, J. J., and Mills, R. H. Study of the reliability of coding of census returns. Paper prepared for September 1962 meetings of the American Statistical Association.

2. Hansen, M. H., Hurwitz, W. N., and Bershad, M. A. Measurement errors in censuses and surveys. Bulletin of the International Statistical Institute, v. XXXVIII, pt. 2, pp. 359-74.

3. Pearson, Karl. On the probability that two independent distributions of frequency are really samples from the same population. Biometrika, v. 8, 1911-12, pp. 250-4.

4. U. S. Bureau of the Census. The Post-Enumeration Survey: 1950. Bureau of the Census Technical Paper No. 4, Washington, D. C., 1960.

ATTACHMENT A

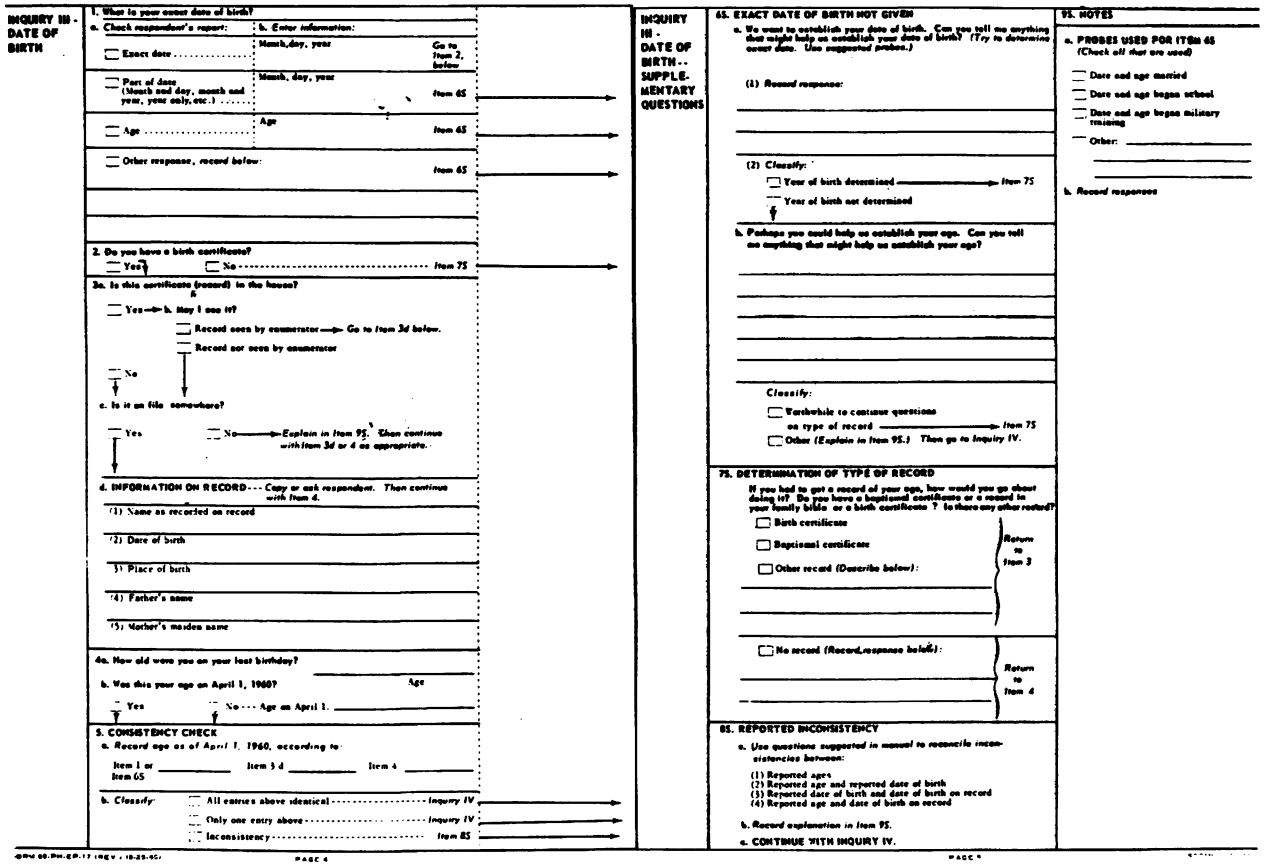


Table 1.--PRELIMINARY ESTIMATES OF THE BIAS IN THE STATISTICS ON LABOR-FORCE STATUS IN THE 1960 AND 1950 CENSUSES OF POPULATION, FOR THE "IDENTICAL POPULATION" FOURTEEN YEARS OLD AND OVER, BY SEX

Sex and labor-force status	1960 Census				1950 Census				Difference between relative biases 2/
	Estimated percentage in class according to		Estimated bias 1/		Estimated percentage in class according to		Estimated bias 1/		
	1960 Census (P_{tG})x100	CPS ($P_{tG'}$)x100 April 1960	Absolute 3/ ($b_{P_{tG}}$)x100	Relative 3/ ($b_{P_{tG}}$) as percent of $P_{tG'}$	1950 Census (P_{tG})x100	CPS ($P_{tG'}$)x100 April 1950	Absolute 3/ ($b_{P_{tG}}$)x100	Relative 3/ ($b_{P_{tG}}$) as percent of $P_{tG'}$	|(8)| - |(4)|
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Males	100.0	100.0	--	--	100.0	100.0	--	--	--
1. In the civilian labor force....	79.1	81.3	-2.2	-2.7	82.1	84.0	-1.8	-2.2	-0.5
a. Employed.....	75.5	77.5	-2.0	-2.6	78.2	79.2	-1.0	-1.2	-1.4
(1) In agriculture.....	7.2	8.3	-1.1	-13.6	11.9	12.5	-0.7	-5.5	-8.1
(2) In nonagricultural industries.....	68.3	69.3	-0.9	-1.3	66.4	66.7	-0.3	-0.4	-0.9
b. Unemployed.....	3.6	3.8	-0.2	-4.4	3.9	4.8	-0.9	-18.1	+13.7
2. Not in the civilian labor force	20.9	18.7	+2.2	+11.8	17.9	16.0	+1.8	+11.5	-0.3
Females	100.0	100.0	--	--	100.0	100.0	--	--	--
1. In the civilian labor force....	35.1	36.1	-1.0	-2.8	29.6	31.8	-2.2	-6.8	+4.0
a. Employed.....	33.3	34.2	-0.8	-2.4	28.3	30.2	-1.9	-6.2	+3.8
(1) In agriculture.....	1.1	1.2	-0.1	-9.1	0.9	1.7	-0.8	-46.4	+37.3
(2) In nonagricultural industries.....	32.2	32.9	-0.7	-2.2	27.4	28.5	-1.1	-3.8	+1.6
b. Unemployed.....	1.8	2.0	-0.2	-10.3	1.3	1.6	-0.3	-18.9	+8.6
2. Not in the civilian labor force	64.9	63.9	+1.0	+1.6	70.4	68.2	+2.2	+3.2	+1.6

1/ Minus sign indicates understatement in Census; plus sign indicates overstatement.
2/ Minus sign indicates larger bias in 1960 Census than in 1950 Census; plus sign indicates larger bias in 1950 Census.
3/ Computed from unrounded figures.

Table 2.—PRELIMINARY ESTIMATES OF THE "INDEX OF INCONSISTENCY" FOR LABOR-FORCE CLASSIFICATIONS IN THE 1960 AND 1950 CENSUSES OF POPULATION, FOR THE "IDENTICAL POPULATION" FOURTEEN YEARS OLD AND OVER, BY SEX

Sex and labor-force status	Index of inconsistency, $\hat{I}_{dG}$		Difference [1]
	1960 Census	1950 Census	(2) - (1)
	(1)	(2)	(3)
Males			
1. In the civilian labor force	.177	.205	+.028
a. Employed	.170	.196	+.026
(1) In agriculture	.224	.144	-.080
(2) In nonagricultural industries	.132	.140	+.008
b. Unemployed	.500	.513	+.013
2. Not in the civilian labor force	.177	.205	+.028
Females			
1. In the civilian labor force	.192	.195	+.003
a. Employed	.175	.180	+.005
(1) In agriculture	.593	.957	+.364
(2) In nonagricultural industries	.156	.145	-.011
b. Unemployed	.720	.751	+.031
2. Not in the civilian labor force	.192	.195	+.003

1/ Minus sign indicates greater unreliability in 1960 Census than in 1950 Census; plus sign indicates greater unreliability in 1950 Census.

Table 3.—PRELIMINARY ESTIMATES OF THE BIAS IN THE STATISTICS AND OF THE "INDEX OF INCONSISTENCY" FOR SEX AND COLOR IN THE 1960 CENSUS OF POPULATION, FOR THE "IDENTICAL POPULATION"

Characteristic and category	Bias, $b_{p_{tG}} \times 100$ [1]	Relative bias [2]	$\hat{I}_{dG}$
	(1)	(2)	(3)
Sex			
Male	+0.2	+0.4	.018
Female	-0.2	-0.4	.018
Color			
White	+0.2	+0.2	.045
Nonwhite	-0.2	-1.7	.045

1/ Minus sign indicates understatement in Census; plus sign indicates overstatement.

2/ Computed from: $[\frac{b_{p_{tG}}}{p_{tG'}}] \times 100$, where $p_{tG'}$ is the estimate from Study EP-10.

Table 4.--PRELIMINARY ESTIMATES OF THE BIAS IN THE STATISTICS AND OF THE "INDEX OF INCONSISTENCY" FOR FIVE-YEAR AGE CLASSES IN THE 1960 AND 1950 CENSUSES OF POPULATION, FOR THE "IDENTICAL POPULATION"

Age class	Bias, $b_{p_{tG}}$ x 100 [1]		Relative bias [2]			Index of inconsistency, $\hat{I}_{dG}$		
	1960 Census	1950 Census	1960 Census	1950 Census	Difference [3] |(4)| - |(3)|	1960 Census	1950 Census	Difference [3] (7) - (6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0-4	+.01	-.18	+.06	-1.63	+1.57	.020	.025	+.005
5-9	+.02	+.08	+.16	+ .92	+ .76	.029	.028	-.001
10-14	+.05	+.01	+.47	+ .11	- .36	.024	.034	+.010
15-19	-.07	+.11	-1.00	+1.64	+ .64	.029	.040	+.011
20-24	-.04	+.02	-.79	+ .26	- .53	.037	.051	+.014
25-29	+.08	--	+1.53	- .03	-1.50	.036	.062	+.026
30-34	-.03	+.04	-.49	+ .48	- .01	.043	.076	+.033
35-39	+.12	+.06	+1.85	+ .78	-1.07	.058	.075	+.017
40-44	+.03	+.09	-.44	+1.38	+ .94	.078	.088	+.010
45-49	-.12	--	-1.85	- .07	-1.78	.071	.101	+.030
50-54	+.03	+.02	+.59	+ .30	- .29	.078	.112	+.034
55-59	+.10	-.16	+2.11	-3.11	+1.00	.063	.103	+.040
60-64	-.10	-.04	-2.77	-1.04	-1.73	.098	.084	-.014
65-69	+.09	-.02	+2.63	- .52	-2.11	.078	.090	+.012
70-74	-.11	--	-.40	+ .12	- .28	.095	.095	--
75 and over	-.05	-.03	-1.80	-1.07	- .73	.032	.051	+.019

1/ Minus sign indicates understatement in Census; plus sign indicates overstatement.

2/ Computed from: $[\frac{b_{p_{tG}}}{p_{tG'}}]$ x 100, where $p_{tG'}$ is the estimate from Study EP-10.

3/ Minus sign indicates higher level of error in 1960 Census than in 1950 Census; plus sign indicates higher level of error in 1950 Census.

Table 5.--PRELIMINARY ESTIMATES OF THE BIAS IN THE STATISTICS AND OF THE "INDEX OF INCONSISTENCY" FOR MOBILITY-STATUS CLASSES IN THE 1960 AND 1950 CENSUSES OF POPULATION, FOR THE "IDENTICAL POPULATION" FIVE YEARS OLD AND OVER

Mobility-status classes [1]	Bias, $b_{p_{tG}}$ x 100 [2]		Relative bias [3]			Index of inconsistency, $\hat{I}_{dG}$		
	1960 Census	1950 Census	1960 Census	1950 Census	Difference [4] |(4)| - |(3)|	1960 Census	1950 Census	Difference [4] (7) - (6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Same house	+1.4	+0.3	+2.6	+0.3	-2.3	.072	.223	+.151
Different house, same county	+0.3	-0.7	+1.2	-5.8	+4.6	.125	.260	+.135
Different county, same State	-0.7	+0.1	-7.6	+2.3	-5.3	.108	.274	+.166
Different State	-0.9	+0.2	-11.2	+8.3	-2.9	.107	.336	+.229
Abroad	-0.2	+0.2	-13.3	+278.3	+265.0	.187	.584	+.397

1/ Residence five years prior to the Census date for the 1960 Census; residence one year prior to the Census date for the 1950 Census.
2/ Minus sign indicates understatement in Census; plus sign indicates overstatement.

3/ Computed from: $\left[\frac{b_{p_{tG}}}{p_{tG'}}\right] \times 100$, where $p_{tG'}$ is the estimate from Study EP-10.

4/ Minus sign indicates higher level of error in 1960 Census than in 1950 Census; plus sign indicates higher level of error in 1950 Census.

Table 6. - PRELIMINARY ESTIMATES OF THE BIAS IN THE STATISTICS AND OF THE "INDEX OF INCONSISTENCY" FOR EDUCATIONAL ATTAINMENT CLASSES IN THE 1960 AND 1950 CENSUSES OF POPULATION, FOR THE "IDENTICAL POPULATION" TWENTY-FIVE YEARS OLD AND OVER

Educational attainment class	Bias, $b_{p_{tG}}$ x 100 [1]		Relative bias [2]			Index of inconsistency, $\hat{I}_{dG}$		
	1960 Census	1950 Census	1960 Census	1950 Census	Difference [3] |(4)| - |(3)|	1960 Census	1950 Census	Difference [3] (7) - (6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
None.....	-0.1	-1.0	- 0.7	-29.0	+28.3	.238	.554	+.316
Elementary, 1-4 years.....	-0.5	+0.3	+ 8.5	+ 3.6	- 4.9	.309	.360	+.051
Elementary, 5-6 years.....	-0.8	-0.6	-11.0	- 5.7	- 5.3	.333	.479	+.146
Elementary, 7 years.....	-0.8	-1.2	-11.1	-14.3	+ 3.2	.399	.604	+.205
Elementary, 8 years.....	+0.7	+1.4	+ 4.6	+ 7.3	+ 2.7	.300	.400	+.100
High school, 1-3 years.....	+0.7	-0.7	+ 3.6	- 4.0	+ 0.4	.240	.375	+.135
High school, 4 years.....	-0.5	+0.3	- 2.0	+ 1.3	- 0.7	.186	.263	+.077
College, 1-3 years.....	+1.0	+1.0	+11.4	+15.0	+ 3.6	.224	.339	+.115
College, 4 or more years.....	+0.2	+0.5	+ 3.1	+ 8.7	+ 5.6	.074	.170	+.096

1/ Minus sign indicates understatement in Census; plus sign indicates overstatement.

2/ Computed from: $[\frac{b_{p_{tG}}}{p_{tG'}}]$ x 100, where $p_{tG'}$ is the estimate from Study EP-10.

3/ Minus sign indicates higher level of error in 1960 Census than in 1950 Census; plus sign indicates higher level of error in 1950 Census.

EVALUATION STUDIES OF THE 1959 CENSUS OF AGRICULTURE

R. Hurley, T. Jabine and D. Larson
Bureau of the Census
Washington, D.C.

A. Content of the Evaluation Program of the 1959 Agriculture Census

After careful consideration of the resources available, the relative importance of needs for data on different types of measurement error, and past experience in the evaluation of agriculture census data, it was concluded that the Evaluation Program for the 1959 Census of Agriculture should consist of two projects:

1. A national Evaluation Survey designed primarily to obtain estimates of bias and its components for selected items.

2. An Enumerator Variability Study, designed to provide estimates of the enumerator contribution to nonsampling variability for selected items.

B. The Evaluation Survey[1]

1. Objectives

The objectives of the Evaluation Survey were as follows:

a. To provide national and regional measures of the accuracy of census results for a restricted number of important items. The items selected were

Number of farms
Total land in farms
Acres of cropland harvested
Acres of corn harvested
Acres of wheat harvested
Acres of cotton harvested
Number of farms classified by size, tenure, economic class and type

b. To provide data concerning factors associated with measurement errors, such as

(1) Cross tabulations of farms by "match status", i.e., missed in the Census, enumerated with errors, enumerated without error, and by farm characteristics likely to be associated with match status, such as size, tenure arrangements, residence of operator, number of separate tracts, etc.

(2) Classification of errors on farms enumerated in the Census by reason for error.

c. To conduct a test of procedures for the measurement of bias for selected livestock items. No livestock items had been included in the Evaluation Survey for the 1954 Census of Agriculture, mostly because satisfactory measurement techniques were not believed to be available for these items. For the 1959 program, it was decided to attempt to develop and test procedures for measuring the bias of:

Cattle and calves on hand
Hogs and pigs on hand
Litters of pigs farrowed, 12/1/58 to 11/30/59

2. Survey Design

The survey design of the Evaluation Survey may be summarized as follows:

a. An area sample was used to obtain information for farms missed in the Census. This sample had two parts -

(1) A rural area sample of 772 segments in 196 primary sampling units, comprising 348 counties. This was a self-weighting sample, with an overall sampling fraction of 1 in 1,500.

(2) An urban area sample, comprising seven-eighths of the segments in urban areas in the November 1959 sample for the Census Bureau's Current Population Survey.

b. A list sample was used to obtain information on places[2] which had been included in the Census. This sample also had two parts.

(1) A sample consisting of all enumerated places of less than 5,000 acres whose locations had been marked by Census enumerators within the boundaries of the 772 segments in the area sample. This sample included about 2,770 farms and nonfarm places.

(2) A sample of 562 Census farms of 5,000 acres and over, selected independently of the area sample.

c. The basic survey operations may be summarized as follows:

Step 1 - An independent enumeration designed to obtain detailed and accurate information for all farms in the area sample and all places operated by persons in the list sample.

Step 2 - Matching of the results obtained in Step 1 against the Census materials in order to tentatively identify farms missed in the Census and errors for places included in the Census.

Step 3 - Followup, as required, to clarify and check the results of Steps 1 and 2.

Step 4 - Final processing and tabulation of the results.

d. Estimates of net error were obtained by combining the results of the area and list samples. The area sample provided estimates of the component of net error arising from farms completely missed in the Census. The list sample provided estimates of the component of net error arising from errors in reporting and (to some extent) processing of data for farms included in the Census. Simple unbiased estimates were used in both cases.

The estimated totals (shown in column (2) of Table 1) were obtained by adding the estimated net error for each item to the corresponding published Census total.

3. Results

The principal results of the Evaluation Survey are shown in Tables 1 and 2. Except for the Census totals, the figures shown are subject to sampling error, and therefore should be interpreted with caution. Sampling errors for the estimates in Table 1 are shown in column (6) of that table. Some of the principal results which may be noted in Table 1 are as follows:

a. The Evaluation Survey estimate of the number of farms in the conterminous United States was 4,045,000, as compared with a Census total of 3,704,000. The difference, or net error, was 341,000, or 8.4 percent of the Survey estimate.

b. The estimates of net underenumeration for the principal acreage items were 6.0 percent of the estimated total for land in farms, 4.3 percent for acres of cropland harvested, 4.5 percent for acres of corn, 3.0 percent for acres of wheat and 9.1 percent for acres of cotton.

c. The relative net errors for acreage items, with the exception of cotton, are significantly smaller than the relative net error for farms.

d. Again with the exception of cotton, the estimates of relative net error for the 1959 and 1954 Censuses are quite similar for those items for which data are available from both years.

e. A considerable part of the net error in number of farms is accounted for by marginal farms. Approximately two-thirds of the net error in number of farms was accounted for by farms of less than 100 acres, and about three-fourths of the net error by farms with total sales of less than $2,500 in 1959.

f. The net errors for livestock and dairy farms were smaller than the net errors for other types of farms, especially those specializing in field crops. This is probably associated with the fact that farms in the latter group have a higher proportion of nonresident operators.

In Table 2, the percent distributions of Evaluation Survey (ES) farms by three match status classifications are shown for several different characteristics of farms and farm operators. The match status classifications used in this table are based on the extent to which these particular farms, as identified in the Evaluation Survey, were covered by the Census enumeration. The basic classification contains two groups - enumerated in Census and missed in Census - and the enumerated in Census group is sub-divided into "complete matches" and "partial matches", this classification depending on the extent to which Evaluation Survey and Census data for the farm differed. A study of these results will reveal several factors that are associated with failure to find and enumerate farms and, once a farm is located, failure to identify correctly the land included in the place.

Some of the kinds of farms that were most commonly missed were farms with nonresident operators (26.3 percent of the ES farms), small farms (24.3 percent of the ES farms under 10 acres and 17.1 percent of the ES farms with 10 to 49 acres), farms in enumeration districts which were not entirely rural in character (26.1 percent) and farms with operators under 25 years of age (22.6 per-

cent). The more separate tracts a farm had, the less likely it was to be missed. Partnership operations were less likely to be missed than individual operations; and farms operated by part owners and managers were less likely to be missed (4.7 percent) than those operated by full owners (12.1 percent), with tenant operated farms occupying an intermediate position (9.1 percent missed).

One of the questions asked of each farm operator interviewed during the part of the Evaluation Survey enumeration which took place after the Census was whether a Census questionnaire had been filled for his place. Of those who answered no to this question, an estimated 44.9 percent had actually been missed in the Census.

Some of the kinds of farms for which the Census enumerator was least likely to obtain correct information on acres in place were farms with land in more than one county (42.6 percent complete match, 48.2 percent partial match), farms with nonresident operators (30.3 percent complete versus 43.4 percent partial), farms of 1,000 acres and over (30.9 percent complete versus 64.2 percent partial), farms operated in partnership (45.2 percent complete versus 51.4 percent partial), farms whose most distant tract was at least 10 miles away from the operator's residence (34.2 percent complete versus 62.0 percent partial), farms with three or more landlords (33.6 percent complete versus 60.8 percent partial), and farms whose operators had three or more tenants (16.7 percent complete versus 75.0 percent partial).

The difficulty of determining total acres correctly clearly increased in proportion to the number of separate tracts in the place, in proportion to the numbers of landlords and tenants and in proportion to the size of the place. Apparently, it was considerably easier for the Census enumerator to determine the acreage of a place operated by a full owner, provided he found it in the first place, than it was to do this for a place operated by a part owner, manager or tenant. Changes in the acres in place which occurred during the Census year (1959) also increased the likelihood that the Census enumerator would get an incorrect figure for total acres.

C. The Enumerator Variability Study

1. Objectives

An experiment to measure the enumerator contribution to the non-sampling variability of census data was carried out in connection with the 1950 U.S. Census of Population. The results of this experiment had far-reaching effects on the design of procedures for the 1960 Census of Population. The 1959 Agriculture Census offered an opportunity to apply these measurement techniques to another kind of data and also to try an experimental design believed to be more efficient than the one used in the 1950 Population Census.

The purpose of the Enumerator Variability Study conducted in conjunction with the 1959 Agriculture Census enumeration was to obtain estimates, for selected items, of the nonsampling variability associated with census enumerators. Enumerator variability is the component of total enumerator error which tends to average out through compensating errors over the work of a large number of enumerators. This source of variability therefore has the greatest relative effect for census statistics which are based on the work of only a few enumerators, such as statistics for an enumeration area (one enumerator) or a county (usually a dozen or fewer enumerators), or, if the item occurs infrequently, for larger areas.

2. Design of the Study

Because the purpose of the study was the evaluation of enumerator variability in an agricultural census, the variability measured was not a "pure" enumerator variability but was the variability associated with actual census procedures. These procedures allowed some self-enumeration by respondents (questionnaires were distributed to farmers in advance of the Census enumeration with requests for completion prior to the enumerator's visit) and the usual editing of questionnaires during the census processing operation. Therefore the enumerator variability that was measured might be thought of as a residual variability still remaining after the possible diminishing effects of the above two factors.

The experiment was restricted to an area consisting of ten contiguous counties in the State of Indiana. Consequently, the results are not directly applicable to a nationwide census of agriculture. The area selected was one in which the enumeration of farms does not present any unusual or atypical problems.

To estimate enumerator variability, an interpenetrating sample design was used. Two weeks prior to the Census, listings were prepared of all the places with specified types of agricultural operations in the ten counties. There were about 17,800 such listings made in 104 enumeration areas (EA's). The listings

for each EA were divided into two random halves, which were then assigned, at random, to the two Census enumerators designated to work in the EA. In this way, the work of each of the two enumerators in an EA provided a basis for unbiased estimates of Census statistics for the EA.

From the resulting data, estimates of enumerator variability for 46 items were made for each EA.

The estimate used for a given EA was based on the difference of two variance components -- a "between" enumerator variance minus a "within" enumerator variance. Since each EA estimate was based on data from only two enumerators, the individual estimates were pooled over the 104 EA's in the ten counties to obtain more reliable overall measures of enumerator variability.

3. Results

Results from this study are presented in Tables 3 and 4, and Chart 1. The principal conclusions which have been drawn so far from the analysis of these results are as follows:

a. Levels of enumerator variability for most of the items studied are sufficiently high so that this factor requires careful consideration in the planning of future censuses and surveys.

b. The number of listings for which questionnaires were not obtained, and the number of nonresponses to particular items showed some of the highest levels of enumerator variability. This finding confirms results from the study of enumerator variability carried out during the 1950 Census of Population.

c. The levels of enumerator variability (expressed as a coefficient of variation) for attributes appear to be, on the average, about one-half of the sampling error for a 25 percent simple-random sample (see Chart 1). This statement is valid only for EA's of the approximate size used in this study (the mean number of listings per EA of places with specified types of agricultural operations was about 170). The experiment has not yielded conclusive information on how enumerator variability would be affected by changes in cluster size.

d. Even with the relatively large samples used in this study, it was not possible to obtain sufficiently reliable estimates of enumerator variability for variables, such as area, production and inventories. A satisfactory evaluation of enumerator variability for these items will require the development of improved sample design and estimation procedures.

FOOTNOTES

[1] Further information on the Evaluation Survey procedures and results may be found in the introduction to Volume II, General Report, for the 1959 Census of Agriculture. A detailed report on methodology, entitled "Checking the Accuracy of Area Statistics Obtained in the United States Censuses of Agriculture" may be had on request by writing to the Statistical Research Division, Bureau of the Census, Washington, 25, D.C.

[2] We use the term "place", rather than "farm", because the list sample represented all places for which questionnaires were filled, whether or not they qualified as farms.

Table 1.--ESTIMATES OF NET ERROR FOR SELECTED ITEMS FOR THE UNITED STATES: 1954 AND 1959 CENSUSES OF AGRICULTURE

(Numbers do not add exactly to totals in all cases due to rounding.)

Item	Estimated total (000)	Census total (000)	Estimated net error		Sampling error of estimated percentage net error
			Amount (000)	Percent of estimated total	
(1)	(2)	(3)	(4)	(5)	(6)
Number of farms.....1959	4,045	3,704	341	8.4	1.2
1954	5,201	4,782	419	8.1	0.9
Farms by size:					
Under 10 acres.....1959	298	241	58	19.3	11.1
1954	588	484	104	17.7	4.1
10 to 49 acres.....1959	890	811	79	8.8	3.8
1954	1,364	1,213	151	11.1	2.2
50 to 99 acres.....1959	745	658	87	11.7	2.8
1954	925	864	61	6.6	1.9
100 to 219 acres.....1959	1,038	998	40	3.9	2.5
1954	1,271	1,210	61	4.8	0.9
220 acres and over.....1959	1,074	997	78	7.2	1.2
1954	1,053	1,011	42	4.0	0.9
Farms by economic class, 1959:					
Class I, II and III (Sales of $10,000 and over) .....number	817	794	23	2.8	1.0
Class IV and V (Sales of $2,500 to $9,999).....number	1,328	1,270	58	4.4	1.0
Class VI and other (Sales of $50 to $2,499) .....number	1,897	1,637	260	13.7	2.5
Farms by tenure, 1959:					
Full owner.....number	2,251	2,116	135	6.0	2.6
Part owners and managers.....number	931	830	101	10.8	2.7
Tenants.....number	863	758	105	12.2	2.5
Commercial farms by type, 1959:					
Cash grain.....number	451	398	53	11.7	2.4
Other field crop.....number	519	469	50	9.6	3.5
Vegetable, fruit and nut.....number	94	82	12	12.9	7.4
Dairy.....number	439	428	10	2.4	1.6
Other livestock, except dairy and poultry.....number	697	684	14	2.0	1.1
General.....number	224	212	12	5.4	2.2
Poultry and miscellaneous commercial.....number	149	140	9	6.3	2.0
Total commercial.....number	2,573	2,413	161	6.2	1.0
Land in farms.....acres 1959	1,191,706	1,120,158	71,548	6.0	0.9
1954	1,223,891	1,158,192	65,699	5.4	1.9
Cropland harvested.....acres 1959	325,110	311,285	13,824	4.3	0.9
1954	346,580	332,870	13,710	4.0	1.1
Corn harvested.....acres 1959	83,396	79,616	3,781	4.5	1.1
1954	80,886	78,123	2,763	3.4	1.2
Wheat harvested.....acres 1959	51,088	49,567	1,521	3.0	1.3
1954	54,263	51,362	2,901	5.3	4.1
Cotton harvested.....acres 1959	16,132	14,649	1,483	9.2	2.7
1954	19,026	18,854	172	0.9	1.5

Table 2: DISTRIBUTION OF EVALUATION SURVEY FARMS IN RURAL AREA SAMPLE BY MATCH STATUS, FOR SELECTED CHARACTERISTICS

Characteristic	Number of farms in sample	Percent of total	Percent distribution by match status			
			Total	Enumerated in Census		Missed in Census
				Complete match	Partial match	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Total Evaluation Survey farms in sample	2,555	100.0	100.0	55.5	34.9	9.6
By age of operator						
Under 25	53	2.1	100.0	49.1	28.3	22.6
25 to 34	283	11.1	100.0	49.8	39.6	10.6
35 to 44	570	22.3	100.0	58.8	34.2	7.0
45 to 54	655	25.6	100.0	57.7	33.9	8.4
55 to 64	543	21.3	100.0	57.8	33.9	8.3
65 and over	394	15.4	100.0	53.0	34.8	12.2
Age unknown	57	2.2	100.0	24.6	47.3	28.1
By type of enumeration district						
A (entirely rural)	1,710	67.0	100.0	58.5	34.8	6.7
B (mostly rural)	665	26.0	100.0	51.0	36.2	12.8
C and D (urban)	180	7.0	100.0	42.8	31.1	26.1
By location of land						
One county only	2,414	94.5	100.0	56.2	34.1	9.7
More than one county	141	5.5	100.0	42.6	48.2	9.2
By residence of operator						
On place	2,258	88.4	100.0	58.8	33.8	7.4
Not on place	297	11.6	100.0	30.3	43.4	26.3
By number of tracts						
1	1,572	61.4	100.0	60.8	27.3	11.9
2	605	23.7	100.0	50.6	43.0	6.4
3	214	8.4	100.0	45.8	46.3	7.9
4	93	3.6	100.0	37.6	61.3	1.1
5 or more	71	2.9	100.0	31.0	66.2	2.8

Table 2: DISTRIBUTION OF EVALUATION SURVEY FARMS IN RURAL AREA SAMPLE BY MATCH STATUS, FOR SELECTED CHARACTERISTICS (cont.)

Characteristic (1)	Number of farms in sample (2)	Percent of total (3)	Percent distribution by match status			
			Total (4)	Enumerated in Census		Missed in Census (7)
				Complete match (5)	Partial match (6)	
By size of farm						
Under 10 acres	148	5.8	100.0	47.3	28.4	24.3
10-49 acres	527	20.6	100.0	51.0	31.9	17.1
50-69 acres	183	7.2	100.0	54.6	31.7	13.7
70-99 acres	278	10.9	100.0	59.0	30.6	10.4
100-139 acres	277	10.8	100.0	60.3	33.6	6.1
140-179 acres	266	10.4	100.0	63.6	30.8	5.6
180-219 acres	171	6.7	100.0	59.7	33.9	6.4
220-259 acres	146	5.7	100.0	59.6	37.0	3.4
260-499 acres	335	13.1	100.0	57.9	39.4	2.7
500-999 acres	143	5.6	100.0	48.9	47.6	3.5
1000 acres and over	81	3.2	100.0	30.9	64.2	4.9
By type of operation						
Individual operator	2,409	94.3	100.0	56.1	33.9	10.0
Partnership	146	5.7	100.0	45.2	51.4	3.4
By tenure						
Full owner	1,374	53.8	100.0	63.3	24.6	12.1
Part owner or manager	621	24.3	100.0	43.1	52.2	4.7
Tenant	560	21.9	100.0	49.8	41.1	9.1
By distance from residence to most distant tract						
One tract only, non-resident, or information on distance not obtained	1,715	67.1	100.0	57.7	30.6	11.7
Less than 1.0 mile	327	12.8	100.0	54.7	38.2	7.0
1.0-4.9 miles	350	13.7	100.0	53.1	42.6	4.3
5.0-9.9 miles	84	3.3	100.0	41.7	53.6	4.8
10.0 miles and over	79	3.1	100.0	34.2	62.0	3.8
By number of landlords						
0	1,375	53.7	100.0	63.3	24.7	12.1
1	809	31.6	100.0	49.3	43.5	7.2
2	228	8.9	100.0	43.9	50.0	6.1
3 or more	143	5.8	100.0	33.6	60.8	5.6

Table 2: DISTRIBUTION OF EVALUATION SURVEY FARMS IN RURAL AREA SAMPLE BY MATCH STATUS, FOR SELECTED CHARACTERISTICS (cont.)

Characteristic	Number of farms in sample	Percent of total	Percent distribution by match status			
			Total	Enumerated in Census		Missed in Census
				Complete match	Partial match	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
By number of tenants						
0	2,171	84.9	100.0	58.6	31.7	9.7
1	275	10.7	100.0	41.5	47.6	10.9
2	73	2.9	100.0	34.2	61.6	4.1
3 or more	36	1.5	100.0	16.7	75.0	8.3
By response to Evaluation Survey question on whether Census questionnaire was obtained for place						
Yes	1,732	67.8	100.0	61.8	35.2	3.0
Don't know	159	6.2	100.0	44.6	34.0	21.4
No	234	9.2	100.0	28.6	26.5	44.9
NA[1]	430	16.8	100.0	48.4	38.8	12.8
By changes in acres in place between Jan. 1 and Dec. 1, 1959						
Same on both dates						
No temporary changes[2]	2,128	83.3	100.0	57.8	32.0	10.2
Some temporary changes[2]	85	3.3	100.0	40.0	50.6	9.4
Larger on Dec. 1	250	9.7	100.0	43.6	49.2	7.2
Smaller on Dec. 1	92	3.7	100.0	46.7	50.0	3.3

[1] The large NA rate was due to the fact that this question could be asked only in the post-Census canvass. Operators in pre-Census segments were revisited only in those areas where livestock items were investigated.

[2] A temporary change occurred when land was rented in or rented out by the operator during 1959, but reverted to its owner prior to December 1, 1959.

Table 3: 1959 CENSUS OF AGRICULTURE - ENUMERATOR VARIABILITY STUDY
ENUMERATOR COEFFICIENT OF VARIATION FOR AN EA FOR
"0-1" OR "0-1-NA" ITEMS

Item	Number of listings having item (1)	Proportion of listings having item P (2)	Enumerator coeff. of var. 1/ v (3)	Standard error of enumerator coefficient of variation 1/ $\hat{\sigma}_v$ (4)	Standard error of enumerator coefficient of variation 1/ $\frac{\hat{\sigma}_v}{v}$ as % (5)
Number of listings:					
1. A-1's taken	16,481	.926	.040	.011	27.5
2. A-1's meeting 1959 definition of a farm	14,725	.827	.033	.008	24.2
3. Refusals and not-at-homes	280	.016	1.021	.209	20.5
Tenure:					
4. Tenants	2,652	.149	.048	.166	345.8
5. NA on tenancy	139	.008	1.027	.204	19.9
Number of A-1's reporting acres of:					
6. Owned land	13,630	.766	.047	.020	41.1
7. Cropland	13,530	.760	.039	.018	44.7
8. Cropland in pasture	8,955	.503	.090	.015	17.0
9. Corn	12,255	.688	.037	.020	54.0
10. Soybeans	6,493	.365	2/	2/	2/
Number of A-1's reporting livestock and livestock production:					
11. Cattle	10,124	.569	.048	.007	14.2
12. Milkcows	6,476	.364	.070	.009	12.4
13. Hogs	8,646	.486	2/	2/	2/
14. Chickens	7,578	.426	.084	.004	5.1
15. Chicken eggs sold	5,613	.315	.110	.019	17.0
Off-farm work and other income:					
16. Having off-farm work	8,511	.478	.070	.007	10.0
17. Having 100 or more days of off-farm work	6,165	.346	.092	.005	5.4
18. Having off-farm income greater than farm sales	6,138	.345	.191	.007	3.7
19. NA for off-farm work	477	.027	1.000	.032	3.2
20. NA for off-farm income greater than farm sales	1,589	.089	.641	.041	6.4

1/ The form of these estimates is given in the Appendix.

2/ Estimate of variance negative.

Table 4: 1959 CENSUS OF AGRICULTURE - ENUMERATOR VARIABILITY STUDY
ENUMERATOR COEFFICIENT OF VARIATION FOR AN EA FOR
QUANTITATIVE ITEMS

Item	Total Value (1)	Average value per listing $\bar{x}$ (2)	Enumerator coeff. of var. 1/ v (3)	Standard Error of enumerator coefficient of variation 1/ $\hat{\sigma}_v$ (4)	Standard Error of enumerator coefficient of variation 1/ $\frac{\hat{\sigma}_v}{v}$ as % (5)
Number of acres:					
1. Acres of owned land.....	1,421,108	79.8	2/	2/	2/
2. Acres in the place.....	2,122,129	119.2	.035	.025	70.8
3. Acres of cropland.....	1,307,603	73.5	.058	.020	34.3
4. Acres of cropland pasture	205,208	11.5	.083	.055	65.8
5. Acres of corn.....	567,627	31.9	.096	.020	20.6
6. Acres of soybeans.....	209,799	11.8	.060	.071	118.5
Crop production:					
7. Bushels of corn.....	32,535,857	1828	.096	.040	41.7
8. Bushels of soybeans.....	4,625,036	260	2/	2/	2/
Livestock and livestock production					
9. Number of cattle.....	251,663	14	2/	2/	2/
10. Number of milk cows.....	74,046	4	.077	.034	43.8
11. Number of hogs.....	550,573	31	.045	.041	91.9
12. Number of chickens.....	2,566,351	144	2/	2/	2/
Total NA and answered-in-error entries: 3/					
13. Answered-in-error.....	1,220	.068	.647	.121	18.7
14. NA's excluding completely refused and not-at-home categories.....	9,698	.545	.640	.018	2.8

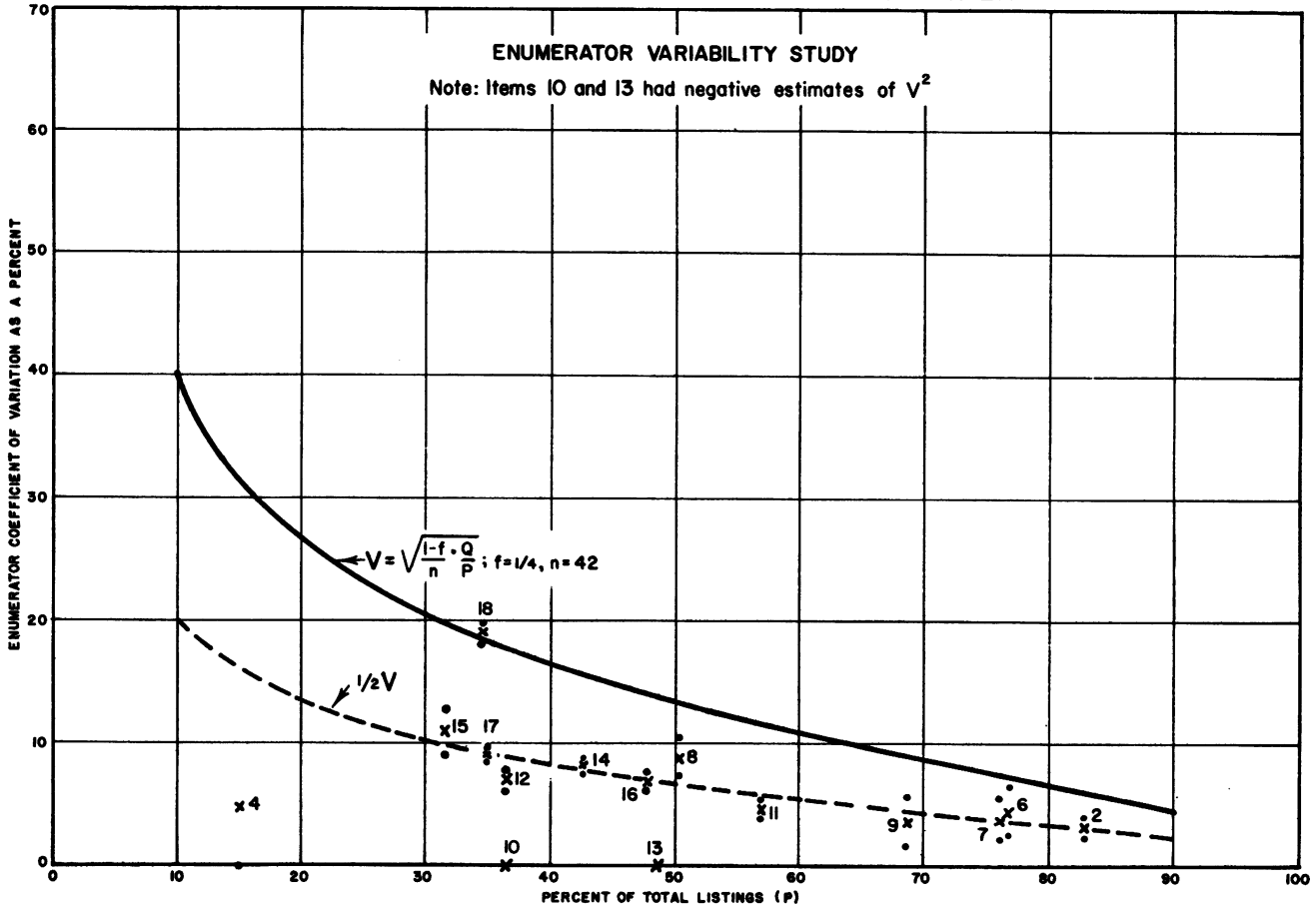
1/ The form of these estimates is given in the Appendix.

2/ Estimate of variance negative.

3/ The original A1 entry was not acceptable according to Census editing rules. These included cases of internal inconsistencies, duplications, extreme values, etc.

CHART I

ENUMERATOR COEFFICIENT OF VARIATION FOR AN EA



APPENDIX:
The Estimate of Enumerator Variability

The Estimate of v^2

A listing of places was first made by means of the normal type of canvass. (It will be noted that the normal type of canvass introduces some geographical stratification into the listing). Within each EA, the listings were divided into groups of ten consecutive listings and each group was randomly divided into two enumerator assignments of size five each.

For the i-th EA (i = 1, 2, ..., 104) within the experiment the average value per listing for a given item would be

$$\bar{x}_i = \frac{\sum_{h=1}^{k} \sum_{c=1}^{b_i} \sum_{j=1}^{\bar{\bar{n}}} x_{ihcj}}{k b_i \bar{\bar{n}}}$$

where $\bar{\bar{n}} = 5$ = the number of listings in each group of ten assigned to one enumerator

b_i = the number of groups of ten in the i-th EA

$k = 2$ = number of enumerators assigned to the i-th EA.

The estimate of enumerator variance for $\bar{x}_i$ is $C_i - D_i$

where
$$C_i = \frac{\sum_{h}^{k} (\bar{x}_{ih} - \bar{x}_i)^2}{k - 1}$$

$$D_i = \frac{\sum_{h}^{k} \sum_{c}^{b_i} \sum_{j}^{\bar{\bar{n}}} (x_{ihcj} - \bar{x}_{ihc})^2}{k b_i^2 \bar{\bar{n}}(\bar{\bar{n}} - 1)}$$

$$\bar{x}_{ih} = \frac{\sum_{c}^{b_i} \sum_{j}^{\bar{\bar{n}}} x_{ihcj}}{b_i \bar{\bar{n}}}$$

$$\bar{x}_{ihc} = \frac{\sum_{j}^{\bar{\bar{n}}} x_{ihcj}}{\bar{\bar{n}}}$$

The estimate of enumerator variance for a total

$$x_i = \sum_{h}^{k} \sum_{c}^{b_i} \sum_{j}^{\bar{\bar{n}}} x_{ihcj}$$

is

$$N_i^2 (C_i - D_i)$$

where $N_i = k b_i \bar{\bar{n}}$ = the number of listings in the i-th EA.

Using information from all EA's within the experiment, an estimate for the relative enumerator variance for an EA total is

$$v^2 = \frac{\frac{1}{L} \sum_{i=1}^{L} N_i^2 (C_i - D_i)}{(\frac{1}{L} \sum_{i=1}^{L} x_i)^2}$$

where $L = 104$ = the number of EA's in the experiment.

The values of v for selected items are recorded in column (3) of Tables 3 and 4.

The Variance of v

To estimate the variance of v, the 104 EA's were divided randomly into three groups of 35, 35, and 34 EA's. For the g-th random group a v_g^2 was computed. Then an estimate for the variance of v would be

$$\hat{\sigma}_v^2 = [\frac{1}{G} \frac{1}{G-1} \sum_{g=1}^{G} (v_g - \bar{v})^2]$$

where $G = 3$.

However for many of the items some of the estimates of v_g^2 were negative so the above estimate, $\hat{\sigma}_v^2$, could not be used. In its place an approximation[1] of the following type was used.

$$V_v^2 \doteq \frac{1}{4} V_{v^2}^2$$

$$\frac{\sigma_v}{v} \doteq \frac{1}{2} \frac{\sigma_{v^2}}{v^2}$$

$$\hat{\vartheta}_v \doteq \frac{1}{2} \frac{\hat{\sigma}_{v^2}}{v}$$

where $\hat{\sigma}_{v^2} = \{\frac{1}{G} [\frac{1}{G-1} \sum_{g=1}^{G} (v_g^2 - \bar{v}^2)^2]\}^{\frac{1}{2}}$

The $\hat{\vartheta}_v$ for each item appears in column(4) of Tables 3 and 4.

Comparison of the Enumerator Variability with Sampling Variability

Chart 1 provides a basis for comparing the enumerator coefficient of variation for the (0, 1) variables with the corresponding coefficient of variation for a 25 percent simple random sample without replacement. For each of the 13 items normally covered[2] in a Census a v is plotted along with the points $v \pm \hat{\vartheta}_v$. The curve which represents one-half of the $\hat{\vartheta}_v$ coefficient of variation for a 25 percent simple random sample from an EA with 168 listings provides a fairly good fit for these items.

FOOTNOTES TO APPENDIX

[1] See Morris H. Hansen, William N. Hurwitz and William G. Madow, "Sample Survey Methods and Theory, Vol. II; Theory," Chap. 10, sec. 1, John Wiley and Sons, Inc., New York, N. Y., 1953.

[2] (0, 1) variables excluded were: (1) Al's taken, (2) Refusals and not-at-homes, (3) NA on tenancy, (4) NA for off-farm work, (5) NA for off-farm income greater than farm sales.

STUDY OF THE RELIABILITY OF CODING OF CENSUS RETURNS

Herman H. Fasteau, J. Jack Ingram and Ruth H. Mills
U. S. Bureau of the Census

General

In operations in which verbal descriptions are coded to alphabetic or numeric equivalents, the codes assigned may have varying degrees of reliability depending upon the quality of the verbal description, the coder's training and experience, and the coding instructions and materials he uses. Customarily, measures of the reliability of this type of coding have been in terms of clerical error rate, i.e., the proportion of cases coded in error.

The purpose of this paper is to outline a technique for evaluating the reliability of individual codes, to show the application of the technique to industry and occupation coding in the 1960 Census of Population, and by an analysis of selected codes to identify some areas in which improvements can be made in questionnaire design, training of respondents and interviewers, code structures, rules for coding, coding materials, and training of coders. Among the reasons for choosing industry and occupation coding for this analysis are: the operation is characterized in part by use of judgment on the part of the coder; there is a tremendous variety of verbal descriptions which must be fitted to a relatively small number of symbolic codes; the operation itself is relatively short term, being preceded by an intensive training program; and, the source materials and procedures are the results of years of research and development with improvements and changes being accumulated from Census to Census. The results contained in this report are partial and preliminary and will be included in a fuller report to be published later.

The Coding Operation[1/]

Information supplemental to basic Census questions was obtained in the Census from a sample consisting of members of every fourth household and from every fourth person not a member of a household. Industry and occupation questions were asked of each person in this sample who had worked since 1950, who was 14 years of age or over and who was not in the Armed Services at the time of the Census. The questions concerned the job held the previous week, or, if none, the last job held. They were:

1. "For whom did he work?
   (Name of company, business, organization or other employer.)"
2. "What kind of business or industry was this?
   (For example: County junior high school, auto assembly plant, TV and radio service, retail supermarket, road construction, farm.)"
3. "Is this primarily Manufacturing, Wholesale trade, Retail trade, or Other?"
4. "What kind of work was he doing?
   (For example: 8th grade English teacher, paint sprayer, repairs TV sets, grocery checker, civil engineer, farmer, farm hand.)"
5. Was this person an employee of (a) a private company, business or individual, for wages, salary, or commissions, (b) a Federal, State, county or local government, was he (c) self-employed in own business, professional practice or farm or was he (d) working without pay in a family business or farm?

About 82 percent of the information for persons in the 25 percent sample was obtained by self-enumeration; the respondent had these five questions before him and wrote out the answers himself. These answers were later transcribed by Census enumerators to the regular Census schedules. The third question did not appear on the Census schedule; however, in transcription, the enumerator was instructed to use this information in the answer to question number 2. Where any information was missing on the self-enumerated schedule, the enumerator was instructed to obtain the missing information from the respondent. In those cases directly enumerated the enumerator was instructed to obtain the answer to question number 3 as though it were included on the schedule and write the answer as a part of question number 2.

Based upon the answers to the above questions, the coder was required: to assign one of the 149 industry codes and one of the 296 occupation codes to the person, to verify that there was a code for class of worker on the schedule, and to determine that all three codes were in correct combination. Codes are stated in single alphabets or in groups of three digits, e.g., A "Agriculture" or 023 "Professional, technical, and kindred workers, clergymen."

The documents moved through the coding units a State at a time. Each coder was supplied with "Company Name Lists" for the counties and Standard Metropolitan Statistical Areas (SMSA's) in the State he was coding. The "Company Name

Lists" included the names and industry codes of all manufacturing companies known to have fifty or more employees and all other types of companies known to have 100 or more employees (obtained from the 1958 Economic Censuses). He was also provided with an Alphabetical Index of Occupations and Industries[2] containing about 30,000 different ways in which the 296 occupations could be described or combined with specific types of industries and somewhat fewer number of ways in which the industries could be described.[3]

The following is a rough summary of the series of steps taken by the coder in assigning the codes:

1. He searched the appropriate Company Name List for the name of the company given in question 1, above.

   a. If the company was found, he immediately assigned the designated industry code. If the company was not found,

   b. He matched the answers to questions 2 and 3 (the kind of business or industry) against the Alphabetical Industry Index. If the answers fitted a description in the Alphabetical Index, he assigned the designated code. If the description was not sufficient (after searching the words out of alphabetical sequence) to arrive at a code from the Industry Index, he used any relevant words in the company name or occupation entry.

   If an appropriate entry were found, he assigned the code given in the Index. If a highly similar entry could not be found, or there was contradictory information in the answers to the four questions, he referred the item to an expert coder.

2. Having assigned the industry code, he then matched the answer to question 4 (occupation description) to the occupation Index in a manner similar to 1(b), above. The occupation codes for a given occupation entry in the Index, however, often vary from one major industrial group or detailed industry code to another and/or by class of worker. If the industry code already decided upon and/or the class of worker on the schedule were not provided as allowable in combination with the occupation code, the item was to be referred to an expert coder.

3. Class of worker was then verified as to its consistency with the industry and occupation codes assigned.

Control of the Quality of Coding

The data used for this study are largely a by-product of the quality control scheme used in the 1960 Census. This scheme is discussed more fully in an earlier paper.[4] For the purposes of this discussion, it can be described briefly as follows:

For industry and occupation coding verification there was a selection of a 1 in 40 sample of households from the 25 percent Census sample. Persons in the experienced civilian labor force falling into the 1 in 40 sample households were coded independently by three different clerks having approximately the same training and coding experience. All three of these coders coded from the Census Schedule, but only the third and last coder entered his code on the Census Schedule. This coder is referred to below as the "Census Coder." The coded results were then matched against one another and placed into one of the four categories:

1. All three agreed on the code (category AAA).

2. Two agreed on a code and one disagreed (category $\overline{AAB}$).

3. Two agreed to refer the item and one coded it (category $\overline{RRA}$).

4. All three disagreed (category $\overline{ABC}$).

For quality control purposes, if two coders agreed and the third disagreed, a quality demerit was assigned to the disagreeing coder. The demerit was utilized as though it represented a defective item. In controlling the quality, categories RRA and ABC were excluded from the sample; however, in this analysis all four categories were included.

Methodology

The sample for this study includes one-fourth of the AAA cases and all cases of disagreement of any kind occurring in the 1 in 40 sample. A measure was developed for scaling the codes in terms of degree of consistency of application. It is called an Index of Consistency[5] and answers the question: "Given a code, how consistently was it independently applied by three different coders looking at the same description?"

The Index takes the form:

$$C = \frac{3\,N_{AAA} + N_{\overline{AAB}}}{3\,N_{AAA} + 2\,N_{\overline{AAB}} + (N_{\overline{ABB}} + N_{\overline{ABC}})}$$

where

N_{AAA} = the number of cases for which all three coders applied the code under consideration.

$N_{\overline{AAB}}$ = the number of cases for which any two of the coders applied the code under consideration and the third applied some other code.

$N_{\overline{ABB}}$ = the number of cases for which only one coder applied the code under consideration and the other two coders agreed upon some other code.

$N_{\overline{ABC}}$ = the number of cases in which all three coders disagreed but one of them applied the code under consideration.

The index of consistency is an average measure of agreement among coders in independently assigning a given code when looking at the description. It is not a direct measure of the quality of the codes in the Census, but serves as a useful device in pointing out those codes which may have low reliability because of confusion on the part of the coder as to what code to assign for a given description. A referral action is here treated as a bona fide code.

An auxiliary measure of consistency answers the question: "Considering the difference cases for a code, what are the associated codes and how frequently are they so associated?" Difference cases are all those cases not falling into the AAA category. An "associated" code is the code or codes that the other two coders entered when one coder entered the code under consideration. The purpose of this question is to point out broad paths which lead to identification of specific areas in which improvements can be made. The answer leads to further questions as to how and why the codes were assigned.

As a starting point, industry codes with low indices of consistency were selected for a two-phase analysis. The first phase of the analysis was to answer the above questions; the second phase was to determine the correct code for the difference case. One part of determining the correct code was to obtain the code which the coder should have assigned using all the information at his disposal; the other part was to obtain the one correct code which was arrived at through the use of detailed research materials coupled with the expert's wide experience. This distinction in "correct" codes is necessary because in some cases the descriptions obtained in the Census could lead to a wrong code, or in some cases the only action that the coder could correctly take was to refer.

The second phase of the analysis called for assigning a 50 percent sample of the difference cases to a panel of experts who assigned the correct code for each case. They received only the answers as obtained in the Census; they had no way of knowing what codes had been assigned by the three coders.

Once the correct code has been assigned, each of the above difference cases was analyzed in terms of:

1. The relation of the code under consideration to its associated code or codes and to the correct code. (It is possible that neither the code under consideration nor an associated code is correct.)
2. The correctness of the code used in the Census tabulations, whether or not the compared codes crossed major Census classifications, and whether or not they were "basket-type" codes such as "miscellaneous," or "not elsewhere classified," etc.
3. The relation of the Census description to entries in the coder's reference material according to ease and type of matching.
4. The presence or absence of "key" words in either the description obtained in the Census or that given in the reference material.
5. The apparent reason for the incorrect code, if any.[6] Major classifications of reasons are: (a) Coder's failure to follow instructions; (b) inadequate descriptions on the Census schedule; (c) inadequate reference materials or instructions; (d) inadequacies in both the description on the schedule and the reference materials or instructions; (e) clear-cut clerical errors such as the transposition of digits.

Indices of Consistency

Table 1 gives the distributions of the 149 industry codes and 296 occupation codes by size of consistency index. A substantially larger proportion of occupation codes has high indices of consistency than do industry codes. Thirty-nine percent of the industry codes as compared with 48 percent of the occupation codes had indices between .90 and .99. For both types of coding, the .90 to .99 class accounted for about 74 percent of the experienced civilian labor force.

Table 1.--Number of Industry and Occupation Codes by Index of Consistency

Index of consistency	Industry codes			Occupation codes		
	Number of codes	Percent of codes	Estimated percent of Labor Force 4/	Number of codes	Percent of codes	Estimated percent of Labor Force 4/
.900-.999	59	39.3	73.6	142	48.0	74.4
.800-.899	56	37.3	18.8	93	31.4	22.5
.700-.799	23	15.3	6.6	41	13.9	2.6
.600-.699	6	4.1	0.8	11	3.7	0.2
.500-.599	2	1.3	0.1	4	1.3	0.3
Less than .500	4	2.7	0.1	5	1.7	3/
Total codes 1/	150	100.0	100.0	296 2/	100.0	100.0

1/ Includes the codes for "not reported."
2/ Excludes Code 000 "Accountants and auditors" because of programming error.
3/ Less than 0.5 percent.
4/ Estimates based on a sample of 420,000. These cases do not include those in which the Census coder referred the description to an expert.

Inspection of the distribution in Table 1 indicated that perhaps the greatest payoff from a preliminary analysis in depth would occur in analyzing the 20 percent or so codes having the lowest indices. The cutoff point for both industry and occupation codes is at Index .80.

The fact that the two above numbers seem to be complementary is only coincidental. Table 2 is a list of the 35 industry codes having Indices less than .80, and the most interesting thing about that list is that it includes every industry code for wholesale trade (Codes 606-629).

Table 2.--List of Industry Codes Having Indices of Consistency of Less Than .800

Code	Description	Number of cases in sample		Index
		AAA	Other	
	Manufacturing, Durable Goods			
208	Miscellaneous wood products, except furniture	76	379	.739
236	Miscellaneous nonmetallic mineral and stone products	129	469	.788
247	Fabricated structural metal products	307	1,200	.775
249	Not specified metal industries	-	38	.073
	Manufacturing, Nondurable Goods			
326	Not specified food industries	26	343	.568
367	Miscellaneous fabricated textile products	11	370	.462
389	Miscellaneous paper and pulp products	-	430	.283
419	Miscellaneous petroleum and coal products	28	138	.739
429	Miscellaneous plastic products	130	446	.799
459	Not specified manufacturing industries	-	94	.060
	Transportation, Communication and Other Public Utilities - Transportation			
516	Warehousing and storage	112	606	.724
519	Petroleum and gasoline pipelines	8	105	.564
526	Services incidental to transportation	52	350	.675
	Utilities and Sanitary Services			
568	Gas and steam supply systems	127	469	.788
569	Electric-gas utilities	103	438	.761
579	Other and not specified utilities	10	54	.714
	Wholesale and Retail Trade - Wholesale			
606	Motor vehicles and equipment	106	423	.775
607	Drugs, chemicals and allied products	87	340	.775
608	Dry goods and apparel	74	326	.760
609	Food and related products	479	1,625	.798
616	Farm products - raw materials	86	472	.725
617	Electrical goods, hardware, and plumbing equipment	208	705	.798
618	Machinery, equipment and supplies	137	1,049	.661
619	Petroleum products	111	654	.710
626	Miscellaneous wholesale trade	552	1,895	.795
629	Not specified wholesale trade	68	583	.642
	- Retail trade			
637	Dairy products stores and milk retailing	51	393	.660
676	Lumber and building materials retailing	364	1,400	.778
687	Fuel and ice dealers	105	475	.753
689	Miscellaneous retail stores	447	1,722	.779
696	Not specified retail trade	85	850	.607
	Business and Repair Services			
807	Miscellaneous business services	597	2,580	.760
809	Miscellaneous repair services	315	1,163	.786
	Professional and Related Services			
888	Nonprofit membership organizations	266	1,014	.781
898	Miscellaneous professional and related services	85	551	.683

Since "Wholesale trade" was the only industry group for which all the codes had low Indices of Consistency and since its codes alone comprised almost 30 percent of the highly confused industry codes, those ten codes were selected for this analysis.

Associated Codes

Often just knowing the codes (and their descriptions) that have a high degree of association with low-consistency codes can simplify the identification of problems and provide clues to the reasons for inconsistency. This is well demonstrated by the wholesale trade codes.

For seven of the ten "Wholesale trade" codes the associated code was "Referral" more frequently than any other code. The codes not included in the seven are "Motor vehicles and Equipment," "Drugs, chemicals and allied products" and "Dry goods and apparel," Next to "Referral" the code most frequently associated with the code under consideration was one specifying the same type of product but in retail trade or manufacturing. Table 3 shows the distribution of general classes of associated codes for each of the Wholesale codes.

Table 3.--Types of Codes Associated With the Wholesale Trade Difference Cases

Wholesale codes	Percent distribution of types of associated codes				Total cases (100%)**
	Referrals	Same type product but retail or Mfg.*	Other wholesale codes	Other codes	
Motor vehicles and equipment	22	49	10	19	(415)
Drugs, chemicals and allied products	18	42	20	20	(328)
Dry goods and apparel	14	45	19	22	(319)
Food and related products	19	57	8	16	(1,587)
Farm products-raw materials	21	7	20	52	(462)
Electrical goods, hardware and plumbing equipment	27	24	14	35	(697)
Machinery, equipment and supplies	34	21	14	31	(1,018)
Petroleum products	30	44	13	13	(634)
Miscellaneous wholesale trade	18	--	18	64	(1,729)
Not specified wholesale trade	45	--	30	25	(566)

*For example, "Manufacturing-motor vehicles and motor vehicle equipment" and "Motor vehicles and accessories retailing" for the first Wholesale trade code listed. See Appendix B for a list of the ways in which the specific codes were grouped.

**The number of cases here is less than the number of difference "Other" cases presented in Table 2 because the cases where the associated code was a blank or an impossible code have been excluded from this table.

It can be assumed that a large proportion of the cases falling into the class "Referrals" and "Same type product but retail or manufacturing" was due to a lack of sufficient information supplied to the coder. To a large extent this is true, as will be shown below. However, "Referral cases are not confined to the "Same type product" class. Appendix C shows how Wholesale codes are associated in coding with the various industry groups. A referred case could have been coded ultimately into any of these industry groups.

Preliminary Analysis of Inconsistent Wholesale Codes

One of the more interesting results of the analysis of these codes is that, if the coder had followed his instructions precisely, he would have referred 75 percent of the difference cases because the descriptions and/or instructions were not adequate for proper coding. In spite of this, in 45 percent of the difference cases the Census coder managed to arrive at the correct code, and in an additional 19 percent of the cases he did refer; so that the presumption is that 64 percent of the difference cases were correctly coded in the Census (assuming that the experts handling the referral cases assigned the correct code). Perhaps one of the reasons for not referring a case on the part of the Census coders was that the rules for referring were very rigid and the coder often felt he knew the correct code.

On an overall basis, for the Wholesale codes, the inconsistent code under analysis was the ultimately correct code for the given case in slightly more than half of the cases. This is not always true, however, for each of the ten codes. At the extremes: when "Wholesale trade, not specified" (629) was used, it was correct in 32 percent of the cases; on the other hand, when "Wholesale trade, electrical goods, hardware and plumbing equipment" (617) was used, it was correct 72 percent of the time.

Table 4.--Codes Under Consideration (Wholesale trade) in Difference Cases, Classified by Correctness and Effect Upon Census

Classification	Code under consideration										
	Total	606	607	608	609	616	617	618	619	626	629
Code under consideration correct:	55%	44%	40%	52%	53%	56%	72%	57%	64%	56%	32%
Assigned by Census coder	32	20	31	40	35	38	36	33	33	32	17
Assigned by another coder but Census coder assigned a different code	12	14	3	6	11	9	18	16	9	12	0
Assigned by another coder but Census coder referred	11	10	6	6	7	9	18	8	22	12	15
Code under consideration incorrect:	45	56	60	48	47	44	28	43	36	44	68
Assigned by Census coder	19	30	28	9	22	27	9	12	13	18	37
Assigned by another coder but Census coder assigned a different code	18	19	26	27	16	12	17	25	13	18	12
Assigned by another coder but Census coder referred	8	7	6	12	9	5	2	6	10	8	19
Total number of cases (100%)	(918)	(59)	(32)	(33)	(221)	(56)	(78)	(141)	(86)	(171)	(41)

From Table 5 it can be seen that a little over half of the Wholesale trade difference cases had adequate descriptions. In 32 percent of the cases the descriptions in the schedule were inadequate for coding; and in 14 percent of the cases the coding materials themselves were inadequate.

Table 5.--Relation of Industry Description on Census Schedule Coding Materials,* Wholesale Trade Difference Cases

	Percent of difference cases		
Adequate Description			54
No coding error made (combination of correct code and referral)		25	
Highly similar words in alphabetical Index	14		
Easy inference required to match to coding materials	11		
Could have been correctly coded		29	
Highly similar words in alphabetical Index	20		
Easy inference required to match to coding materials	9		
Inadequate Description			32
Problems of detailed classification**		19	
Description more general than alphabetical Index	11		
Description more detailed than alphabetical Index	6		
Other	2		
Problems of major classification only		13	
Coding Materials Inadequate			14
Total difference cases (100%)			(918)

*"Coding materials" refers to the alphabetical Index and/or the "Company Name List."
**In some of these cases there can also be a problem of major group classification.

In 25 percent of these Wholesale trade difference codes none of the three coders assigned an incorrect code -- one or two of them assigned the correct wholesale code and two or one of them referred the case.

In 29 percent of the difference cases, one of the three coders assigned an incorrect code when there was no excuse for it. If the coders had followed their instructions explicitly, they would have arrived at the correct code.

In 19 percent of the cases the industry description on the schedule could not be properly matched to the Index for detailed classification (other than major industrial classification). In most of these cases the description on the Census schedule was too general: for instance, entered on the schedule would be "shipping" whereas one needed to know the product shipped or the shipping product manufactured in order to match the description to the Index. Only in half as many cases was the description on the schedule more detailed than the Index; most frequently separate parts of the description could be matched to different lines in the Index leading to different codes.

In 13 percent of the difference cases there was no trouble in matching the description to the Index, but information as to the major industrial classification was absent or incorrect.

In the remaining 14 percent of the difference cases the coding materials themselves were inadequate. In a sizeable proportion of these cases the "Company Name Lists" were in error.

Some Suggested Actions for Improvement

While the above observations apply primarily to codes in wholesale trade, they indicate some actions which can be taken to improve the reliability of industry coding. The first of these has to do with improvement of the description of industry activity provided to the coder. The emphasis would be upon the respondents and would utilize Census public relations media in informing them how the question should be answered and the importance of answering it correctly. An accompanying action would be intensification of training and control of interviewers in requiring and obtaining correct answers to the industry activity question.

It was noted earlier that the major industrial classification was not explicitly provided for on the Census schedule. It was asked explicitly on the self-enumeration questionnaire. It is possible that the description provided to the coder will be improved if the question is included on the Census schedule. In considering a proposal such as this, however, it is necessary to consider whether dis-economies arising from such a change will be more than offset by improvement in coder reliability. This is a question which must be investigated.

Further Research

This has been a preliminary report presented for the purpose of indicating the procedure in a study of coding reliability. The analysis in depth of wholesale trade difference cases, above, is only a part of that analysis with future emphasis to be placed upon the effect of a lack of reliability upon detailed published Census figures.

In addition to the above, the study will continue largely as follows:

a. Complete the analysis outlined above for all industry codes having low Indices of Consistency.

b. Subject all occupation codes having low Indices of Consistency to similar analysis.

c. Analyze a sample of difference cases, both industry and occupation, for codes having high Indices of Consistency.

d. Analyze a sample of AAA cases to determine how frequently consistency of response led to an incorrect code. Research to date has indicated that at least for occupation codes the incidence of error in AAA cases is extremely small.

e. Investigate the correctness of codes applied in the Census for cases in which all three coders referred.

f. As a result of the above analyses, provide a comprehensive list of changes which can be made to improve the reliability of coding.

FOOTNOTES

1/ The industry and occupation coding specifications and training materials were developed by the Occupation and Industry Section of the Economic Statistics Branch of the Population Division of the Bureau of the Census. Members of this Branch, in particular William J. Mulligan, Stanley Greene and Mrs. Gladys M. Dodd, have cooperated in this analysis of coding consistency.

2/ U. S. Bureau of the Census, 1960 Census of Population, Alphabetical Index of Occupations and Industries, Washington, D. C., February 1960. (Revised Edition, October 1960.)

3/ For example: There are codes for about 90 differently described automobile company entries; there are codes for over 300 different types of occupation descriptions containing the word "engineer" although there were only ten different codes into which engineers can be classified. There are listed about 280 different types of college teachers, 280 different types of "inspectors," 220 "repairmen," 140 "mechanics." There are also such single-entry items as "Krippendorfer" (coded as an "operative and kindred worker, not elsewhere classified" in the "Leather and leather products: footwear, except rubber" industry); "Osmosis man" (coded also as an "operative and kindred worker, n.e.c." but in the miscellaneous food preparation and kindred products industry); and "Shill" (coded as an "attendant, recreation and amusement" in the miscellaneous entertainment and recreation services).

4/ Cf. M. H. Hansen, H. H. Fasteau, J. J. Ingram and G. Minton, "Quality Control in the 1960 Censuses," New Frontiers in Administrative and Engineering Quality Control, ASQC, Milwaukee, 1962, pp. 323-339.

5/ Max Bershad of the Statistical Research Division of the Bureau of the Census developed the model for this Index. See Appendix A for a fuller presentation. This is similar to the Index presented at this meeting by L. Pritzker, R. Hanson, "Measurement of Errors in the Censuses of Population and Housing."

6/ If one coder assigned the code under consideration which turned out to be correct, and the other two coders referred the case because the description was ambiguous, the case was treated as though there were no incorrect code assigned.

APPENDIX A: Derivation of the Index of Consistency

Let: j = the document

$j = 1, 2, 3 \ldots\ldots N$

i = the coder

$i = 1, 2, 3 \ldots\ldots K$

i' = a coder other than the one selected as i

Let: X_{ij} = 1 when coder i classifies the j^{th} document as code h.

= 0 when coder i classifies the j^{th} document as other than code h.

$$X.. = \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} X_{ij}}{KN}$$

For perfect consistency in the use of code h,

$$E(X_{ij} - X..)^2 = E(X_{ij} - X..)(X_{i'j} - X..)$$

$$\text{or } 1 = \frac{E(X_{ij}-X..)\,(X_{i'j}-X..)}{E(X_{ij}-X..)^2}$$

and a measure of consistency for code h would be:

$$C = \frac{E(X_{ij}-X..)\,(X_{i'j}-X..)}{E(X_{ij}-X..)^2}$$

$$= \frac{E(X_{ij}\,X_{i'j} - X..^2}{E(X_{ij}^2) - X..^2}$$

Since X.. in this study will be very small, the measure of consistency becomes

$$C = \frac{E(X_{ij}\,X_{i'j})}{E(X_{ij})^2} = \frac{E(X_{ij}\,X_{i'j})}{E(X_{ij})} = \frac{E(X_{ij}\,X_{i'j})}{X..}$$

The numerator and denominator of C will be estimated from a sample of n documents and of coders, each sample document having been coded independently by three different people.

For three coders, an estimate of $E(X_{ij}\,X_{i'j})$ for code h is

$$\frac{\sum_{j=1}^{n} (x_{1j}x_{2j} + x_{1j}x_{3j} + x_{2j}x_{3j})}{3n}$$

For code h = A, and other codes designated by B,

$$\frac{\sum_{j=1}^{n}(x_{1,j}x_{2,j} + x_{1,j}x_{3,j} + x_{2,j}x_{3,j})}{3n} = \frac{(n_{AAA} + n_{AAB}) + (n_{AAA} + n_{ABA}) + (n_{AAA} + n_{BAA})}{3n}$$

$$= \frac{3n_{AAA} + n_{AAB} + n_{ABA} + n_{BAA}}{3n}$$

$$= \frac{3n_{AAA} + n_{\overline{AAB}}}{3n}$$

An estimate of X.. is

$$\bar{x} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{3} x_{ij}}{3n} = \frac{3n_{AAA} + 2n_{\overline{AAB}} + n_{\overline{ABB}} + n_{\overline{ABC}}}{3n}$$

where C is any code other than A or B.

Therefore, an estimate of

C is $$\frac{3n_{AAA} + n_{\overline{AAB}}}{3n_{AAA} + 2n_{\overline{AAB}} + n_{\overline{ABB}} + n_{\overline{ABC}}}$$

APPENDIX B: Wholesale Trade Codes and Associated Same Product Manufacturing and Retail Trade Codes

Wholesale Code		Associated Codes of Same Type Product		
Code	Description	Code	Description	Percent
606	Motor vehicles and equipments	267	Manufacturing, durable goods, transportation equipment, motor vehicles and motor vehicle equipment	15
		656	Retail trade - motor vehicles and accessories retailing	34
607	Drugs, chemicals and allied products		Manufacturing, non-durable goods, chemicals and allied products	
		406	- Synthetic fibers	-
		407	- Drugs and medicine	10
		408	- Paints, varnishes and related products	4
		409	- Miscellaneous chemicals and allied products	19
		658	Retail trade, drug stores	9
608	Dry goods and apparel		Manufacturing, non-durable goods,	
		349	- Yarn, thread and fabric mills	17
		B	- Apparel and other accessories	21
		367	- Miscellaneous fabricated textile products	*
		646	Retail trade, apparel and accessories stores except shoe stores	7
609	Food and related products		Manufacturing, non-durable goods, Food and kindred products	
		306	- Meat products	14
		307	- Dairy products	10
		308	- Canning and preserving fruits, vegetables and seafoods	10
		309	- Grainmill products	1
		316	- Bakery products	2
		317	- Confectionery and related products	1
		318	- Beverage industries	3
		319	- Miscellaneous food preparation and kindred	2
		326	- Not specified food industries	2
		F	Retail trade - Food stores, except dairy products	12
616	Farm Products - Raw Materials	A	Agriculture	7
617	Electrical goods, hardware and plumbing		Manufacturing, durable goods	
		259	- Electrical machinery, equipment and supplies	16
		649	Retail trade - Household appliances TV, and radio stores	8
618	Machinery, equipment and supplies		Manufacturing, durable goods, Machinery, except electrical	
		256	- Farm machinery and equipment	2
		257	- Office, computing and accounting machines	5
		M	- Miscellaneous machinery	10
		666	Retail trade - Hardware and farm equipment stores	4
619	Petroleum products		Manufacturing, non-durable goods	
		416	- Petroleum refining	24
		419	- Miscellaneous petroleum and coal products	*
		657	Retail trade - gasoline service stations	9
		687	- Fuel and ice dealers	11

*Less than 0.5 percent.

APPENDIX C: Distribution of Codes Associated With Each Wholesale Code by Major Industrial Classification

	606	607	608	609	616	617	618	619	626	629
Referrals	.22	.18	.14	.19	.21	.27	.34	.30	.18	.45
Agriculture	*	-	-	.05	.07	-	*	-	.04	-
Forest and Fisheries	-	-	-	*	-	-	-	-	-	-
Mining	-	*	-	*	*	-	.01	.04	.01	-
Construction	-	*	-	*	-	.03	.01	-	.02	*
Manufacturing	.24	.39	.45	.45	.13	.25	.26	.26	.28	.06
Durable goods	.20	.04	.01	*	*	.24	.25	.01	.14	.02
Nondurable goods	.04	.35	.44	.45	.13	.01	.01	.25	.14	.04
Transp. Commun. and other Public Utilities	.04	.02	.03	.06	.16	.03	.01	.05	.03	.05
Wholesale and Retail Trade	.46	.37	.34	.23	.34	.37	.28	.34	.40	.37
Wholesale	.10	.20	.19	.08	.20	.14	.14	.13	.18	.30
Retail	.36	.17	.15	.15	.14	.23	.14	.21	.22	.07
Finance, Ins. and Real Estate	.01	*	.01	.01	.04	*	*	*	*	.01
Business and Repair Services	.03	.01	.01	.01	.02	.04	.07	.01	.02	.03
Personal Services	-	*	.01	-	*	.01	*	*	.01	*
Entertainment and Recreation Services	-	-	-	*	-	-	*	-	*	*
Professional and Related Services	-	*	-	*	.02	*	.01	*	*	.01
Public Administration	-	*	-	*	*	*	-	-	-	.01
Industry not reported	-	-	*	-	*	-	*	*	*	*
Total **(1.00)	(415)	(328)	(319)	(1,587)	(462)	(697)	(1,018)	(634)	(1,729)	(566)

*Less than .005.

**These totals are slightly less than the totals of the difference cases presented in Table 2 because excluded here are cases where the "associated code" was a blank or an impossible code.

DISCUSSION

Ansley J. Coale, Office of Population Research, Princeton University

The basic problem of evaluating the coverage of a population census is one of finding or generating reliable independent data against which the census can be checked. There are two categories of comparison which may be applied:

1. Aggregate comparisons
2. Individual or name-by-name comparisons

In an aggregate comparison an independent source is found that purports to cover the whole of the population or the whole of some portion that is delineated within the census. If the independent source can be accepted as correct the aggregate comparison indicates how fully the census has covered the whole population or the delineated sub-group. If the independent source can not be accepted as standard the comparison shows the relative completeness of the two.

Three examples of aggregate comparison are:

1. The population under 25 can be compared with estimated survivors of the registered births from 1935 to 1960, adjusted for underregistration of births and net immigration. Such a comparison has been carried out by Akers, of the Census Bureau. The results show an estimated undercount of persons under 25 of 2.8 per cent (white males 2.6 per cent, white females 1.6 per cent, non-white males 8.3 per cent, non-white females 6.2 per cent)

2. In 1940 selective service registration of males of military age could be compared with the census enumeration of this group from six months earlier with due allowance for mortality and differences in coverage. The results showed a 4.5 per cent undercount of white males and an 18 per cent undercount of non-white males at ages 21-35.

3. The population in a given census can be compared with an aggregate obtained by updating adjusted earlier census records. Dr. Melvin Zelnik and I have recently completed such a comparison, which is fully reported in a book now in the hands of the publishers. The results show an undercount for white males of 2.6 per cent and for white females of 1.6 per cent.

The independent data utilized in these examples, especially the first, are of fairly convincing reliability, but it remains less than certain that they provide a wholly trustworthy independent determination of the number of persons.

In an individual or name-by-name comparison the independent source again covers the whole or some delineated portion of the population. In this case the comparison is not one of total figures but is an individual match of persons in the two sources. The advantage of such an individual comparison is that even an incomplete source can indicate the completeness of coverage of the census provided that the chance of omission from the source is independent from the chance of omission in the census. Thus if one had a truly independent list of 10,000 persons who should have been covered by and census and found that 9,700 individuals had been so covered he could reasonably conclude that the census was 97 per cent complete, even though the 10,000 names were not a complete coverage of any well defined sub-group. Examples of individual comparisons are:

1. A comparison of registered births in the three months prior to the census with the enumeration of children no more than three months of age in 1950. This comparison indicated that about 4 per cent of these babies had been omitted.

2. Re-enumerative sample surveys. The result in 1960, reported to us today, indicates a 1.6 per cent net omission.

3. Reverse record checks where the independent source is a list of persons constructed from birth certificates, persons enumerated in earlier censuses, immigration records and the like. Before such a list can be utilized, however, it is necessary, outside of the census records, to locate the 1960 address of each person on the list.

Individual comparisons provide a valid estimate of census coverage only in the absence of correlation between the chance of omission from the census and the independent source. Unfortunately in re-enumerative surveys the correlation is probably very high rather than nearly zero. The chances are very strong that a person omitted from a more intensive re-enumerative survey, for example, persons with no usual place of residence or persons whom the respondent is omitting because of possible difficulties with the law, will be missed in both the census and the survey.

In summary, it is very difficult to achieve complete coverage in a census or to improve the 97-98 per cent completeness already attained. It is also very difficult to determine precisely the extent of coverage when it is incomplete. As the chairman has remarked, the Bureau of the Census is constantly experimenting to improve coverage and, as the papers under discussion demonstrate, is pioneering in the evaluation of its own success. Nonetheless, I would say that as of the moment we do not yet know how complete the 1960 census count was. If we do find out when the evaluation program is completed, my private guess is an undercount of about 3 per cent, compared to some 3.5 per cent in 1950.

V

MEASUREMENT TECHNIQUES IN PUBLIC HEALTH

Chairman, Jacob E. Bearman, University of Minnesota

Page

Quality of Data Elicited by Successive Mailings in Mail Surveys - Monroe G. Sirken and Morton L. Brown, National Center for Health Statistics..... 118

Some Statistical Features of the Health Examination Survey - Walt R. Simmons, Arthur J. McDowell and Tavia Gordon, National Center for Health Statistics..... 126

Measurement of Water Quality Through a National Sampling Network - Gordon E. McCallum, U. S. Public Health Service..... 138

QUALITY OF DATA ELICITED BY SUCCESSIVE MAILINGS IN MAIL SURVEYS

By: Monroe G. Sirken and Morton L. Brown
National Center for Health Statistics
National Vital Statistics Division

Introduction

It is customary in mail surveys to conduct additional mailings with nonrespondents in order to increase the number of respondents and thereby decrease the potential impact of bias error due to nonresponse. Typically, the gains resulting from each additional mailing are measured in terms of the increment to the response rate.[1] In effect, it is assumed that the quality of reporting in the survey is essentially the same regardless of whether the first mailing or one of the additional mailings elicits the reply. Findings based on a mail survey conducted by the National Vital Statistics Division do not bear out this assumption. On the contrary, the adequacy of reporting in this survey was highest for survey questionnaires elicited by the first mailing and was successively lower for questionnaires elicited by each subsequent mailing. This finding, if generally applicable, would suggest that the costs of additional mailings should be justified on the basis of marginal gains in the proportion of adequate responses rather than solely on the basis of reductions in the nonresponse rate.

Survey Design

Recently, the National Vital Statistics Division conducted a survey for the Population Research and Training Center, University of Chicago, involving a probability sample of 9,541 persons who died during the 4-month period, May through August 1960. The 9,541 decedents were selected from the Current Mortality Sample—a ten-percent sample of death certificates which the 54 independent Registration Areas in the Nation transmit monthly to the National Vital Statistics Division.

The mortality sample survey was one phase of a large-scale project to match the death certificates for a sample of approximately 340,000 deaths out of 536,000 occurring in the 4-month period, May through August 1960, with the enumeration records for these decedents in the 1960 Census. "It [the project] is designed to provide nationwide statistics on mortality differentials by the full range of social and economic characteristics collected in the 1960 Census of Population and thus to circumvent the restrictions imposed on mortality analysis by the limited information reported in the official death record."[2] It was anticipated that the matching operation would be unsuccessful in locating a census record for about one-fifth of the decedents. The primary purpose of the mortality survey is to provide "Census" information for a sample of decedents not matched with the 1960 census records in the large-scale matching operation. However, the mortality survey was conducted in advance of the matching operation for a sample of all decedents included in the larger project, since it was considered inadvisable to delay the survey until the fall of 1962—the earliest date by which the "unmatched" decedents could be identified.

Data collection in the mail survey was started in the summer of 1960 and was completed less than one year later. The survey was undertaken as soon as feasible after the decedent was selected into the sample in order to avoid nonresponse losses due to migration of the respondent and also to reduce the risk of response error due to memory loss. Nevertheless, there was an average delay of about 6 months between the date of death and the date that the respondent replied to the mail survey.

The death certificate informant—the person who provided the funeral director with information about the decedent for the death certificate—was the principal respondent in the mortality survey. (Henceforth, we will refer to this person as the "informant.") The name and address of the informant, who is usually a close relative of the decedent, appears on the death certificate. The first questionnaire was sent to informants by regular mail, and nonrespondents were sent 2 additional mailings at 2 week intervals—the first by certified mail, and the second, a special nonresponse letter, by regular mail. Finally, arrangements were made with the Bureau of the Census to conduct personal interviews with nonrespondents in the mail survey who resided in counties included in the sample design of the Current Population Survey and in adjacent counties.

Questionnaires returned in the mail survey were reviewed and evaluated in terms of the "adequacy" of the reported information. In essence, adequacy is a measure of codeability. The rules for measuring adequacy were based entirely on the completeness and internal consistency of the information reported on the mail questionnaire. Obviously, the concept "adequacy" in this sense is not necessarily a measure of validity, since an adequate response may or may not be a valid response. On the other hand, an inadequate response can hardly be a valid response, and for this reason, the adequacy measure would appear to have utility as an index of the quality of response, particularly in those surveys where the absence of a criterion source makes it impossible to validate the reported information.

The absence of a criterion source was not, however, the justification for assessing the "adequacy" of responses in the mortality survey, since eventually the information as reported in

the 1960 census will be available—at least for the "matched" decedents. The measure of adequacy was a by-product of the editing procedure in the mortality survey where it was used in deciding whether or not follow-up mailing actions were indicated in order to improve the quality of reporting on the "original" questionnaire. (By the "original" questionnaire we mean a questionnaire that was elicited by the first, second, or third mailing as distinguished from forms that were elicited by special mailing actions that were undertaken subsequently to improve the quality of reporting.)

The questionnaire covered most of the items contained either in the 100-percent or in the 25-percent sample schedule of the 1960 Census of Population. Specified items in each part of the questionnaire were considered essential (Chart I). If, for any of these essential items, the information was missing or incomplete, or was inconsistent with information reported elsewhere on the questionnaire, the part of the questionnaire containing this item was assessed to be not adequate. The adequacy of each of the 5 parts of the questionnaire was assessed independently and the sum of the adequate parts (ranging from 0 to 5) was the overall measure of the questionnaire's adequacy.

In order to improve the quality of information reported in the mail survey, special mailing actions were frequently undertaken for cases in which the original questionnaire was evaluated as not adequate. The special mailing actions involved either mailing a questionnaire to a referral respondent identified by the informant or mailing a special letter to the informant enclosing a form containing only those questions which were not adequately answered on the original questionnaire. The basic rules for conducting special mailing actions were: (1) a questionnaire was mailed to a referral whenever one was given, or (2) a special letter was mailed to the informant if one or at most two parts of the original questionnaire were not adequate. Prior experience contraindicated conducting follow-up actions routinely with informants whose reports were generally inadequate throughout the original questionnaire.

At the discretion of the director of the mortality survey, however, special letters were occasionally mailed to informants even though more than 2 parts of the original questionnaire were not adequate. On the other hand, follow-up actions with informants who adequately completed most of the parts of the original questionnaires were occasionally contraindicated by public rela-

CHART I - ESSENTIAL ITEMS OF INFORMATION ON EACH PART OF THE QUESTIONNAIRE

PART OF QUESTIONNAIRE	ESSENTIAL ITEMS
I: PLACES OF RESIDENCE OF THE DECEDENT.	PLACE OF RESIDENCE AS OF APRIL 1, 1960 (STREET ADDRESS, CITY OR TOWN, AND STATE).
II: HOUSEHOLD IN WHICH DECEDENT WAS A MEMBER.	NAMES, DATES OF BIRTH, SEX, RACE, MARITAL STATUS, AND RELATIONSHIP TO HEAD OF HOUSEHOLD OF EACH PERSON LIVING WITH THE DECEDENT ON APRIL 1, 1960.
III: CHARACTERISTICS OF THE DECEDENT.	DATE HE LAST WORKED AT A JOB. EMPLOYMENT STATUS DURING LAST WEEK OF MARCH 1960. OCCUPATION AT WHICH LAST WORKED. INDUSTRY IN WHICH LAST WORKED.
IV: CHARACTERISTICS OF DECEDENT'S SPOUSE (OR PARENT IF DECEDENT WAS A CHILD).	DATE HE LAST WORKED AT A JOB. EMPLOYMENT STATUS DURING LAST WEEK OF MARCH 1960. OCCUPATION AT WHICH HE LAST WORKED. INDUSTRY IN WHICH HE LAST WORKED. (NOTE: THESE ITEMS REQUIRED ONLY FOR HUSBANDS OF DECEASED MARRIED WOMEN AND FOR PARENTS OF DECEASED UNMARRIED CHILDREN 17 OR UNDER.)
V: INCOME CHARACTERISTICS OF DECEDENT AND HIS FAMILY.	INCOME OF THE DECEDENT FROM WAGES OR SALARY, FROM PROFITS AND FEES, OR FROM ANY OTHER SOURCE. INCOME OF DECEDENT'S SPOUSE (OR PARENT) FROM WAGES OR SALARY, FROM PROFITS AND FEES, OR FROM ANY OTHER SOURCE. INCOME OF OTHER RELATIVES LIVING WITH THE DECEDENT FROM WAGES OR SALARY, FROM PROFITS AND FEES, OR FROM ANY OTHER SOURCE.

tions considerations or on evidence that the informant did not know the missing information.

After replies were received from the special follow-up mailings, the adequacy of the combined information reported on the original questionnaire and on the follow-up forms was re-evaluated.

Response and Adequacy Rates

About 45 percent of the informants replied to the first mailing and the response rate was raised to 83 and 88 percent, respectively, after the second and third mailings had been completed. The response rate after completion of personal interviews was 94 percent. (Personal interviews were completed with about 82 percent of the 691 informants with whom they were attempted. Personal interviews were not attempted, however, with 496 respondents to the mail survey representing principally those who lived in counties not covered by the interviewers.) Considerable care was exercised in establishing which of the three mailing actions elicited the reply. For example, when necessary, postmarks on returned envelopes were used in order to establish which mailing had elicited the reply.

Not quite 50 percent of the questionnaires returned in the mail survey were completely adequate. For the remainder, the information was not adequately reported in one or more parts of the questionnaire. Special mailing actions were instituted for about three-fifths of the original questionnaires that were evaluated to be not entirely adequate. A questionnaire was mailed to 296 referrals identified by the informants and special letters were mailed to 2,404 informants. Subsequently, the cases were re-evaluated based on the combined information reported on the original questionnaire and on the report forms returned as a result of the special mailing actions. These special actions were successful in improving the adequacy of reporting in the survey (Table 1). The proportion of cases replying to the mail survey for which completely adequate information was obtained increased from about one-half to about two-thirds.

Special follow-up mailing actions improved the adequacy of reporting for every part of the questionnaire (Table 2). Parts of the questionnaire for which information was relatively less well reported on the original questionnaire had the largest absolute improvement. Thus, the proportion of cases for which income was reported adequately increased from 63 to 72 or about 9 percentage points, and the adequacy of residence reporting, which was originally 97 percent, improved by less than 1 percentage point. Nevertheless, the five parts of the questionnaire are ranked by percent adequate in the same order before and after the special mailing actions were conducted. The part of the questionnaire pertaining to the place of residence of the decedent on April 1 was most often adequately completed and the part pertaining to the 1959 income of the decedent and his family was least often adequately completed.

The overall adequacy of reporting in this mail survey is not typical of mortality surveys that have been conducted in recent years by the National Vital Statistics Division.[3] By comparison, the informant response rates to this mail survey were slightly lower, and the adequacy of the reported information was poorer. All mortality surveys have a distinct methodological problem; namely, the respondent reports not-for-self. There were additional factors, particular to this survey, pertaining to the items of information collected which probably depressed the adequacy of the derived statistics. The survey questionnaire was lengthy and relatively complex since every attempt was made to duplicate the concepts employed in the 1960 Census. With few exceptions the census question wording and categories of response were adhered to strictly. Thus, April 1, the Census date, was the reference date in many of the questions. This was a somewhat arbitrary, if not an artificial date of reference when applied to persons who died in the four-month period, May through August, and about whom the mortality survey collected information approximately 6 months later. The problem presented by adhering to the Census reference date was even more difficult for such items of information as income and weeks worked for which the reference period was 1959.

TABLE 1. NUMBER OF PARTS OF THE QUESTIONNAIRE FOR WHICH ADEQUATE INFORMATION WAS REPORTED BEFORE AND AFTER SPECIAL FOLLOW-UP MAILINGS WERE CONDUCTED

Number of Adequate Parts	Proportion of Questionnaires That Were Adequate	
	Prior to Follow-up Mailings	After Follow-up Mailings
5	47.6%	65.7%
4	29.1	17.6
3	14.2	8.7
2	5.5	5.1
1	1.6	1.5
0	2.0	1.4

TABLE 2. ADEQUACY OF THE INFORMATION REPORTED IN EACH PART OF THE QUESTIONNAIRE BEFORE AND AFTER THE SPECIAL MAILING ACTIONS

Part of Questionnaire	Percentage of Questionnaires Returned Adequate		
	Prior to Follow-up Actions	After Follow-up Actions	Absolute Change
Part I - Residence of decedent----------	96.5%	97.3%	0.8%
Part II - Household composition----------	78.7	86.5	7.8
Part III - Characteristics of decedent--------	81.5	88.2	6.7
Part IV - Characteristics of related person---	89.9	92.9	3.0
Part V - Income----------	62.8	71.8	9.0

The adequacy of reporting in the mail survey would have been considerably higher except for the income items which represented by far the least adequately completed part of the questionnaire. Excluding the income items, the reported information was entirely adequate for 64 percent of the returned original questionnaires and this was increased to 78 percent after the special follow-up mailing actions were conducted.

Results by Successive Mailings

The procedures for evaluating the adequacy of the information reported on the original questionnaires were independent of the mailing wave that elicited the reply. There were, nevertheless, significant differences in the adequacy of the reported information on the original questionnaire according to the wave that elicited the reply (Table 3). Questionnaires elicited by the first mailing had relatively the highest adequacy level of reporting, and those elicited by the third mailing had the lowest adequacy level. The proportion of questionnaires elicited by the first, second, and third mailings in which the reported information was entirely adequate was 50 percent, 45 percent, and 40 percent, respectively. Although the special mailing actions improved the adequacy of the information reported on the original questionnaires elicited by each of the mailings, the special action was most effective in improving the information reported on questionnaires that had been elicited by the first mailing, and the gains were successively smaller for information reported on questionnaires that had been elicited by the second and third mailings. Consequently, the disparity in the adequacy of information reported on the questionnaires elicited by the first, second, and third mailings was increased as a result of the special mailing actions.

The difference between the mailing waves in the adequacy of the reported information was observed for each of the five separate parts of the questionnaire (Table 4). The difference between mailing waves was greatest for Part V of the questionnaire which was the least adequate part of the questionnaire and the difference was least for Part I which was the most adequate part of the questionnaire. Thus, for Part V, the difference in the percentage of adequate reports between questionnaires elicited by the first and third mailing waves was 12 percentage points; for Part I, the difference was 2 percentage points. For virtually every part of the questionnaire, the special mailing actions produced the largest absolute gains for questionnaires returned to the first mailing and these actions produced the smallest absolute gains for questionnaires returned to the third mailing. Consequently, the disparity between the mailings on the adequacy of every part of the questionnaire was greater after than before the special mailing actions were conducted.

Three factors help to explain the differential effectiveness of the special follow-up mailing actions in improving responses to the questionnaires elicited by the first, second, and third mailings: (1) differences in the proportion of inadequate original questionnaires for which special mailing actions were undertaken, (2) differences in the response rates to the special mailing action, and (3) differences in the adequacy of reporting on the forms returned to the special mailing action. For each of these factors, the results favored the questionnaires elicited by the first mailing most and those elicited by the third mailing least.

A special mailing action was initiated with 64 percent, 60 percent, and 56 percent, respectively, of the informants whose questionnaires were elicited by the first, second, and third mailing. The questionnaires elicited by the later mailings were more likely to contain expressions of disinterest or of lack of information, conditions under which special mailing actions were contraindicated.

There was a differential response rate to the special follow-up mailing actions according to the mailing that elicited the original questionnaire. Thus, the response rates to the special follow-up mailing were 79 percent, 70 percent, and 46 percent, respectively, according to whether the original questionnaire had been elicited by the first, second, or third mailing.

Finally, the adequacy of reporting on forms elicited by the special mailing action favored

TABLE 3. NUMBER OF PARTS OF THE QUESTIONNAIRE FOR WHICH ADEQUATE INFORMATION WAS REPORTED BY THE MAILING WAVE THAT ELICITED THE ORIGINAL QUESTIONNAIRE

Number of Adequate Parts	Proportion of Questionnaires That Were Adequate		
	First Mailing	Second Mailing	Third Mailing
Prior to special follow-up mailing			
5	50.1%	45.5%	40.3%
4	28.9	29.3	28.7
3	12.6	15.3	19.7
2	5.0	6.1	6.4
1	1.7	1.5	1.9
0	1.7	2.3	3.0
After special follow-up mailing			
5	68.9%	62.5%	51.4%
4	15.9	19.0	22.5
3	7.0	9.8	16.7
2	4.8	5.6	5.1
1	1.5	1.5	1.5
0	1.1	1.6	2.8

TABLE 4. ADEQUACY OF INFORMATION REPORTED IN EACH PART OF THE QUESTIONNAIRE BY WAVE ELICITING THE ORIGINAL RESPONSE BEFORE AND AFTER THE SPECIAL MAILING

Part of Questionnaire	Proportion of Returned Questionnaires That Were Adequate		
	First Mailing	Second Mailing	Third Mailing
Prior to special mailing			
Part I - Residence of decedent	96.8%	96.3%	94.4%
Part II - Household composition	80.1	77.6	74.7
Part III - Characteristics of decedent	81.9	81.3	77.9
Part IV - Characteristics of related person-	90.9	88.8	88.8
Part V - Income	65.7	60.3	54.0
After the special mailing			
Part I - Residence of decedent	97.7%	97.1%	94.9%
Part II - Household composition	88.5	84.8	80.5
Part III - Characteristics of decedent	89.1	87.5	84.6
Part IV - Characteristics of related person-	94.0	91.9	90.4
Part V - Income	75.3	69.4	58.0

cases where the original questionnaires were elicited by the earlier mailings. For example, of the inadequate questionnaires for which replies to the special mailing actions were returned, 64 percent became adequate if the original questionnaire was elicited by the first mailing, whereas 52 percent and 33 percent became adequate if the original questionnaire had been elicited by the second and third mailings, respectively. Stated in terms of the proportion of inadequate questionnaires which were not improved at all by the special follow-up action, the results were 30 percent, 47 percent, and 58 percent, respectively, depending on whether the original questionnaire had been elicited by the first, second, or third mailing.

Discussion

We have no simple explanation for the observation that the level of adequacy of the reported information was highest for questionnaires elicited by the first mailing and that it was successively lower for questionnaires elicited by each additional mailing. It seems likely to us that the respondents' ability to answer the questions and their motivation to reply are interrelated factors influencing both the mailing wave to which they reply and the adequacy of the information which they report.

We suspect that the general relationship between the mailing wave eliciting the response and the adequacy of response observed in this survey may be typical of most mail surveys. Preliminary findings derived from this survey indicate that this relationship appears to hold for demographic subgroupings of the decedents. We were particularly curious to investigate this matter for white and nonwhite decedents since earlier mortality surveys [4] had established the fact that response rates were consistently higher for white decedents than those for nonwhite decedents, and in particular, that the response rate to the first mailing was substantially greater for white than for nonwhite decedents. The expected differential response rate by color was also observed in this survey (Table 5). On the basis of the difference in the pattern of response to the 3 mailings we hypothesized that the adequacy of the reported information would be higher for whites than for nonwhites. The statistics substantiated the hypothesis. Fifty-three percent of the original questionnaires returned for white decedents were adequate as compared to 28 percent of the questionnaires returned for nonwhite decedents.

Within each grouping of the decedents by color, the adequacy of reported information was highest for questionnaires elicited by the first mailing and it was successively lower for questionnaires elicited by the second and third mailings (Table 6). We had, however, incorrectly anticipated that the adequacy level for white and nonwhite decedents would be about the same for questionnaires elicited by the same mailing. As it turned out, the information was much more adequately reported for white decedents than for nonwhite decedents for questionnaires elicited by each mailing. It is noteworthy that the adequacy of reporting on questionnaires for white decedents elicited by the third mailing is superior to that for nonwhite decedents elicited by the first mailing. Thus, it appears that the difference in the adequacy of the reported information in the mail survey for white and nonwhite decedents is a function of at least two parameters, the nonresponse rate as well as the proportion of replies elicited by each mailing action.

Summary and Conclusions

In a mail survey involving the collection of household composition and socioeconomic and related demographic characteristics for a national sample of decedents, the initial response rate from relatives of decedents was almost doubled from 45 to 88 percent by means of conducting 2 additional mailings with nonrespondents. The adequacy (i.e., the codeability) of the reported information on the returned questionnaires was substantially improved as a result of mailing special queries to respondents in order to obtain data which originally were not reported adequately. Thus, the proportion of decedents for whom completely adequate reports were obtained increased from about one-half to about two-thirds as a result of the information reported in response to the special mailing actions.

The adequacy of the information reported in the mail survey was dependent upon the mailing action that elicited the original reply. Informants who replied to the first mailing reported most adequately and those who replied to the third mailing reported least adequately. The special mailing actions were relatively more successful in improving the adequacy of information reported on questionnaires elicited by earlier rather than later mailings. Consequently, the special mailings had the effect of increasing the variation in the adequacy of reporting among the mailing waves that elicited the original reply.

There were substantial differences among the 5 parts of the questionnaires in the proportion of returned questionnaires for which the information was adequately reported. Information on income was least adequately reported; for only about 72 percent of the decedents was this information reported adequately after the special mailing actions had been completed. By contrast,

TABLE 5. RESPONSE RATES BY COLOR OF DECEDENT AND BY MAILING WAVE THAT ELICITED THE RESPONSE

Action Eliciting Response	Response Rate		
	Total	White	Nonwhite
Number of decedents	9,541	8,254	1,287
Response to mail survey	87.8%	88.8%	81.7%
First mailing	45.5	46.5	39.2
Second mailing	37.4	37.5	36.9
Third mailing	4.9	4.8	5.6

TABLE 6. NUMBER OF PARTS OF THE QUESTIONNAIRE FOR WHICH ADEQUATE INFORMATION WAS REPORTED BY COLOR OF DECEDENT AND MAILING WAVE THAT ELICITED THE ORIGINAL QUESTIONNAIRE

Number of Adequate Parts of Questionnaire	Proportion of Questionnaires That Were Adequate					
	First Mailing		Second Mailing		Third Mailing	
	White	Nonwhite	White	Nonwhite	White	Nonwhite
Number of decedents-	3,835	504	3,096	475	395	72
Prior to special follow-up mailing						
5----------	52.9%	28.4%	48.6%	25.0%	43.8%	20.8%
4----------	28.9	29.4	29.1	30.7	28.9	27.8
3----------	11.2	23.0	14.3	21.7	18.2	27.8
2----------	4.0	12.5	4.8	14.3	4.6	16.7
1----------	1.4	3.8	1.0	4.8	1.5	4.2
0----------	1.5	3.0	2.1	3.4	3.0	2.7
After special follow-up mailing						
5----------	73.2%	44.0%	65.9%	40.3%	55.2%	30.5%
4----------	15.0	22.4	18.7	20.8	22.0	25.0
3----------	5.7	17.1	8.5	18.3	15.2	25.0
2----------	3.9	11.3	4.3	13.9	3.5	13.9
1----------	1.2	3.6	1.1	4.2	1.3	2.8
0----------	1.0	1.6	1.5	2.5	2.8	2.8

adequate information pertaining to place of decedent's residence was reported for 97 percent of the decedents. Regardless of the mailing wave that elicited the original reply, the 5 parts of the questionnaires were ranked in the same order on the basis of the proportion of questionnaires for which the information was adequately reported. Differences in adequacy of reporting among the 3 sets of questionnaires elicited by the first, second, and third mailings, respectively, were least for the most adequate part of the questionnaire and were increasingly larger for each of the less adequately reported questionnaire parts.

The relationship between mailing wave that elicited the reply and the adequacy of the reported information was observed for both white and nonwhite decedents. These two groups of decedents were investigated because they exhibited different patterns of response (proportion replying to each mailing wave) and different response rates (total responding to the mail survey). Substantial differences were detected between the color groups in the adequacy of the reported information—being better for white than for nonwhite decedents. This difference was partially explicable in terms of the proportions of the returned questionnaires that were elicited by the first, second, and third mailing, respectively; that is the average number of mailings that was required to elicit the original reply was smaller for white than for nonwhite decedents. In addition, however, the adequacy of reporting on questionnaires elicited by each of the three mailing waves was substantially superior for the white decedents who also had the better response rate to the mail survey.

No simple explanation is proposed to account for the relationship observed in this survey between adequacy of the reported information and the mailing wave that elicited the original reply. It is doubtful that the relationship is unique to this survey even though the information collected here was complex and detailed and therefore produced a relatively low level of adequate reporting. Perhaps, the respondents' ability to answer the questions and their motivation to reply are interrelated factors influencing both the mailing wave to which they reply and the adequacy of the information which they report. From this viewpoint, it seems likely that the observed relationship between adequacy of response and the mailing wave that elicited the reply would apply to all mortality surveys conducted by the National Vital Statistics Division and perhaps to mail surveys generally. The finding implies a reconsideration of the model for allocating resources in mail surveys so that the costs of additional mailings and of conducting follow-up interviews are justified on the basis of gains in the proportion of adequate responses rather than solely on the basis of reducing nonresponse.

REFERENCES

[1] Hanson, M. H. and Hurwitz, W. N., "The Problem of Non-Response in Sample Surveys," Journal of the American Statistical Association, Vol. 41, (December 1946) 517-529.

[2] Kitagawa, E. M. and Hauser, P. M., "Methods Used in a Current Study of Social and Economic Differentials in Mortality," The Milbank Memorial Fund Quarterly forthcoming.

[3] Sirken, M. G., Pifer, J. W. and Brown, M. L., "Design of Surveys Linked to Death Records," U. S. Department of Health, Education, and Welfare, National Center for Health Statistics (September 1962).

[4] Sirken, M. G., Pifer, J. W. and Brown, M. L., "Survey Procedures for Supplementing Mortality Statistics," American Journal of Public Health, Vol. 50, No. 11 (November 1960) 1753-1764.

SOME STATISTICAL FEATURES OF THE HEALTH EXAMINATION SURVEY

Walt R. Simmons, Arthur J. McDowell, and Tavia Gordon*

1. The U. S. National Health Survey

The Health Examination Survey is one of three major vehicles being utilized by the U. S. National Health Survey in its program to provide comprehensive data on the amount and distribution of illness, injury and disability of the civilian population of the United States, on other health characteristics of that population, and on the use of medical, dental and hospital facilities. The other two vehicles are the Household Interview Survey, and the Health Records Surveys. All three activities are continuing projects. They are intended to be flexible instruments, complementing one another in providing an intelligence system on what is possibly the nation's greatest resource: the health of its population. The Bureau of the Census has participated in many aspects of the planning, sample selection, and data collection for these Surveys.

The interview survey is particularly effective in assembling data that have their intersection in the subject person himself, and concern medical and health matters of which the person has knowledge; for example, his instances of physician contact, his days of disability, his medical costs. The interview survey is not a suitable mechanism for estimating volume of undetected or undiagnosed diseases. Operating now for five years, it has been described elsewhere in some detail.[1,2,3]

The Health Records Surveys are themselves a family of undertakings characterized by two attributes: (1) The point of contact and the initial sampling unit is the facility or source which provides health care--the hospital, nursing home, personal care place, or the physician's or dentist's office; (2) the data themselves come in large part from records in these places. The records surveys of the U. S. National Health Surveys are in their early stages, and will not be further described here, except to say that they are expected to provide a wide range of information on use of health facilities and on diagnostic detail not readily available through other techniques.

*National Center for Health Statistics, Washington, D. C.

2. The Health Examination Survey

The Health Examination Survey (HES) is a process which collects health data in two essential steps: (1) a probability sample is drawn for some major sector of the national population--e.g., all noninstitutional civilian adults in the age range 18-79 years; (2) persons in this sample are given a limited physical examination, and other relevant measurements are taken, using standardized procedures. The primary objectives also are two in number: (1) to provide basic distributions of the population by a variety of physical and physiological characteristics such as height, weight, blood-pressure, body-build, and visual acuity; and (2) to estimate prevalence of specified chronic conditions. The HES is a continuing activity, which comes in parts that have been termed cycles. A given cycle concerns a particular segment of the national population and a particular set of measurements and conditions. The remainder of the present discussion will be focused on the first cycle of the HES, for which collection of data will be completed in December of this year.

3. Statistical Problems

The statistical problems of the Health Examination Survey are numerous and complicated. We can touch on only a few of them. Beyond the critical matters of general objectives, authorization, financing, capital resources, and administrative affairs, the HES exhibits substantial problems in formal survey design, in determination of unit costs, in determining initially speculated population parameters, in choice of estimating equations, in handling the nonresponse issue, in calculating precision of estimates, in training and supervision of the examining staff, in the logistics of field operation, and perhaps most notably in evaluation and control of the many measurement procedures.

The scope and content of the first cycle survey, and a number of these statistical problems have been treated in a publication of the National Health Survey.[4] Discussion in this paper is restricted to necessarily brief accounts of three problem areas in the first cycle: impact of nonresponse on the probability design, measure-

ment problems, and basic estimating equations. In all three instances, the final story cannot yet be told since study continues on each of the topics.

4. The Impact of Nonresponse

The Health Examination Survey, like other parts of the National Health Survey, is based on legislation which specifies that the required information will be secured "on a noncompulsory basis."[4] Thus, from the outset, much attention was given to the problem of nonresponse, since a poor response rate could prevent any valid generalizations of survey findings to the population sampled. We attempted, first, to study the nature and the dimensions of the nonresponse problem, next, to take all practicable steps in the design and conduct of the survey to minimize the extent of nonresponse, and, finally, to obtain auxillary data on both respondents and nonrespondents in order to facilitate the residual imputational process.

Several surveys involving health examinations had been made in local areas in the early 1950's.[5,6,7] That experience indicated that perhaps something like one-third of the people asked to participate in a survey involving examination might fail to cooperate despite intensive persuasion efforts. In preparation for the Health Examination Survey, the National Health Survey undertook methodological studies into the motivations and attitudes involving willingness to participate in a health examination survey.[8,9]

We will make no attempt to list all of the ways in which the survey design incorporated lessons learned from the methodological studies, from earlier related surveys, or from the pilot testing of HES plans, in order to maximize response. It would not be accurate to suggest that the operating program adopted and continued every one of the characteristics which these studies implied might help the response rate. Frequently operational considerations overrode theoretical indications. Frequently, too, the practice of the art of obtaining cooperation in a health examination survey produced techniques which seemed effective. These were used without controlled experiments to determine their contribution to the desired result. Nevertheless, it is possible to identify some of the measures adopted because the methodological studies, the pilot tests, or the earlier surveys suggested they might minimize the amount of nonresponse.

Among a number of factors which seemed to be related to willingness to participate was the potential examinee's knowledge about the nature of the examination and the purpose of the survey. In order to increase confidence in the survey and to allay the fears and doubts that might stem from lack of knowledge about just what was involved, several steps were taken. These included distributing in advance in the sample neighborhoods leaflets containing information about the Health Examination Survey. The sample person who made an appointment was given another leaflet which described in some detail the specific steps that would be carried out in examining him. Also, support of local groups and officials was sought and this and the medical research goals were stressed in stories in local news media.

Another factor thought to be relevant to response was the potential personal benefit to the examinee from his examination. To take advantage of this it was decided that the findings of the examination would be made available to the examinee's personal physician (or, in the case of dental findings, to his dentist) if the examinee would so instruct us.

It had been established that a possible negative motivational factor might be the inconvenience, in travel and loss of time and other ways, which the examination entailed. We attempted to minimize that by restricting the length of the examination, eliminating some possibly embarrassing procedures, careful scheduling to avoid waiting, selecting convenient locations, providing transportation, and other means.

Aside from the various motivational findings, there were other results of the presurvey studies and the pretests which influenced our survey design in the effort to maximize response. Thus, the studies had indicated that persons were less likely to agree to an examination on behalf of another member of the household than they were to agree to come themselves; consequently, we made it a practice to ask about consent to a health examination only directly of the sample person.

For another thing, it appeared important to identify the probable noncooperator early, and to handle his case in an individualized way. Motivations differ widely. Thus, to one person, the fact that our examination is free is a point in its favor; to another, perhaps concerned about government budgets, this may be a negative factor. The possibility of early detection of a disease, if it is present, is welcomed by some and dreaded by others. The plan of the survey called for the initial interviewer to make no further effort to obtain cooperation from a person once he was identified as an apparent noncooperator. Another representative of the survey, armed with all the knowledge collected by the first and with experience in handling problems of this sort, would

make a later call and attempt to explain the survey more fully, correct misapprehensions and obtain cooperation.

How much the success of the Health Examination Survey owes to any or all of the many specific actions that have been taken to maximize response is not a question we can answer now, nor perhaps even later. But we have been encouraged by the survey results so far. As of the time this is written (early July 1962) the Health Examination Survey has completed operations at more than three-fourths of the separate locations throughout the United States that constitute the first cycle sample (33 of 42). The total number of sample persons identified in those areas was 6,105 and we have succeeded in examining 5,235--or 86 percent -- of them. Thus, the rate of nonresponse to date (14 percent) is only about one-half as big as we feared it might be.

The rate referred to combines all reasons for nonresponse--refusal to cooperate, unavailability during the period of the survey for reasons other than health (e.g., away on vacation), or unavailability for reasons related to health (e.g., in short-term hospital). The rate given above relates the total number of sample persons identified in the sample households interviewed and the number of such persons examined. Because the sample is based on households, including a small number for which we are unable to obtain interviews, a further correction for presumed sample persons in noninterviewed sample households is appropriate. This correction would lower the response rate by about two percentage points, from 86 percent to 84 percent. One other measure which needs mention is the average of the percentages of response at the 33 locations. Computed this way, the uncorrected response rate is 86.6 percent.

Tables 1 and 2 show how the response rates have varied with population density group and with geographic region. As expected from the preliminary studies, the response rate varies inversely with population concentration. In rural and other urban (less than 50,000 inhabitants) areas, more than 90 percent of the sample population was examined. This decreases steadily as the concentration of population increases, until in the giant metropolitan areas only 78 percent of the sample was examined. The geographic grouping shows a fairly clear picture of lower response rates in the northeastern part of the United States as compared with the south or the west. For the total region this effect is, of course, enhanced by the relatively great number of giant metropolitan areas in the northeast but the relationship is observable in each group.

So far in this paper we have dealt with our concern about and study of the problem of nonresponse, the operating measures taken to minimize nonresponse and the general results to date. In addition, it is, of course, necessary to make some evaluation of the extent to which the residual nonresponse group differs qualitatively from those who were examined. Considerable information bearing on this results from the fact that the design of the survey incorporates a household interview for each sample household. This interview gives information similar to that collected in the Health Interview Survey. We have this for 98 percent of the HES sample households. In addition to such demographic variables

Table 1. Average Response Rate by Population Concentration Group and by Geographic Region: First 33 Stands, Health Examination Survey

Population Concentration Group	All Regions	Northeast	West	South
All Groups----------	86.6%	81.3%	87.9%	89.9%
Giant Metropolitan Areas----------	77.9	77.2	79.0	-
Other Very Large Metropolitan Areas----------	85.3	-	86.5	84.0
Other SMSA's----------	88.3	85.0	91.0	87.7
Other Urban----------	90.8	89.0	94	91.0
Rural----------	91.7	79	92.7	95.0

NOTE: The figure shown for each cell represents the unweighted arithmetic mean of the response percentages for each of the stands included in that category; thus, the marginal total figures cannot be derived directly from the values in the particular row or column. The two cross-classifications designated by a dash (–) are ones for which no stands have been included; the two for which whole-number percentages are shown are each based on only one stand.

Table 2. Frequency Distribution of First 33 Stands by Population Concentration Group and Percentage of Sample Persons Examined (Response Rate), Health Examination Survey

Population Concentration Groups	All Stands	Percentages of Sample Persons Examined				
		94-100	87-93	80-86	73-79	66-72
All groups	33	6	12	7	6	2
Giant Metropolitan Areas	8			2	4	2
Other Very Large Metropolitan Areas	4		2	1	1	
Other SMSA's	8		6	2		
Other Urban	6	3	1	2		
Rural	7	3	3		1	

as age, sex, race, education, and income group, this provides a health history of the household member, including hospitalization, chronic disease, days absent from work due to illness, etc. Table 3 compares the response and nonresponse groups on the basis of age and sex. While the older age groups are slightly underrepresented in the first Round, there is marked agreement between the distributions for examined persons and for all sample persons.

The Health Examination Survey includes one other important means of obtaining relevant data on not-examined persons. At the time the household interview is completed, an attempt is made to obtain a signature on the medical authorization form, giving us permission to request information from the family physician's medical records. Thus, for most not-examined persons some information directly relevant to the health characteristics which the survey is attempting to measure can be obtained. This is done for all not-examined persons for whom a medical authorization was obtained and in the remaining instances the not-examined person is asked to forward to his physician a similar request. In addition, for comparison, similar inquiries are sent to the physicians of a matched sample of examined persons.

5. Measurement in the HES

Every examination has something unique about it which could affect its comparability with other examinations. For example, the Health Examination Survey uses a drink of 50 grams of glucose in its glucose tolerance test, and this could yield results incomparable with results from a challenge of 100 grams. The examination includes a venipuncture and an electrocardiogram and these procedures could affect the blood pressure of the examinee. At every point the question can be raised of the comparability of the Health Examination measurements with those obtained by other examinations using different instruments or measuring the characteristics in different contexts. Needless to say, the immense number of such points of possible incomparability requires that a choice be made of those factors which seem, on a priori grounds, to be most important to evaluate and most accessible. It is easy, of course, to make the wrong choices: the history of science is littered with such errors.

Such issues of comparability are critical in any study. They determine what we might call the exterior significance of the data. Internally, the problems of standardization are similar, but more easily dealt with. Essentially, they require a protocol, actions to assure conformity, and procedures which permit evaluation of residual variability. The Health Examination Survey has expended considerable efforts in this direction.

An important means of standardizing a medical examination--obvious enough but not always thought of in this connection--is the choice of examining physician. By choosing physicians of similar background and experience (almost all of the examining physicians for the Health Examination Survey are third or fourth year residents in internal medicine) it is reasonable to expect a uniformity of result that would otherwise be obtained only by prohibitively extensive training.

Having devoted reasonable efforts to developing a sound protocol, to choosing, training, retraining and supervising the examining staff, and attempting to control the quality of data collection by a variety of means, we are left with a

Table 3. Unweighted Total Sample Persons and Examined Persons by Age and Sex: First-Round of First Cycle, Health Examination Survey

Age Group	Total Sample			Examined Persons		
	Total	Male	Female	Total	Male	Female
	Percentage distribution by age groups					
Total	100.0	100.0	100.0	100.0	100.0	100.0
18-24	13.1	12.9	13.3	13.2	13.3	13.1
25-34	21.5	21.3	21.6	22.1	21.9	22.3
35-44	22.8	23.0	22.7	23.7	23.6	23.8
45-54	18.2	17.9	18.6	18.4	17.7	19.0
55-64	14.0	15.1	13.1	13.2	14.1	12.5
65-74	8.2	8.1	8.3	7.8	8.1	7.5
75-79	2.1	1.7	2.5	1.6	1.4	1.8
	Percentage distribution by sex					
Total	100.0	45.6	54.4	100.0	46.7	53.3
	Number of persons					
Total	2,614	1,191	1,423	2,205	1,030	1,175

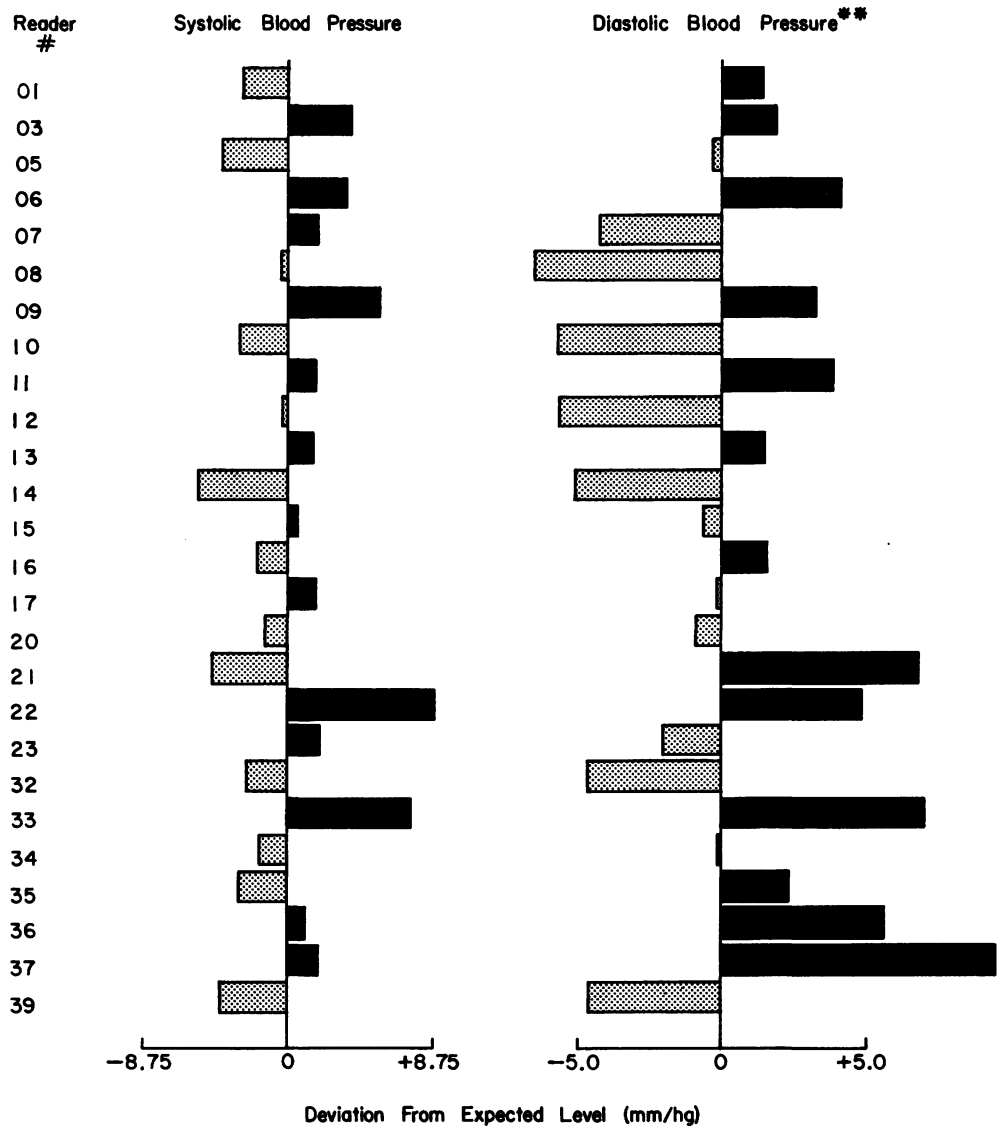
final requirement: determining the variability of measurement. This is something especially attractive to statisticians; indeed, they sometimes seem more interested in quantifying variability than in limiting it. But even limited, some variability will remain. It may be more or it may be less. It may be measurable in the context of the procedures used, or it may not. But unattended it can devour all significance.

The problems of identifying and measuring nonsampling variability arising from sources such as the examining physicians, X-ray readers, and the like, are complex. Even given that reasonable definitions and a good statistical model can be constructed to reflect the effect of nonsampling sources of error, the Health Examination Survey presents formidable administrative problems in attempting to carry out the measurement processes in such a way as to provide measures of nonsampling variability.

Ideally, what would be desired would be replicate measurements of the same characteristics on randomized samples of persons. For characteristics delineated entirely in the course of the clinic examination, this has thus far proved too difficult to arrange. In order to try to approximate this type of evaluation it was arranged, however, to have two examining physicians at most locations, and examinees at a location were assigned alternatively to each physician. Analyses of differences in physical examination findings between physicians indicate, as might be expected, that there is often a greater variability between physicians than chance assignment of examinees would be expected to yield. A good example is blood pressure (Figure A). Data for each physician, on which the chart is based, have been adjusted or standardized for age-sex composition. The exhibited deviations contain components of chance, geography, and perhaps other factors as well as inter-physician differentials, but observed differences between physicians are significant for both systolic and diastolic data. As would be anticipated on an a priori basis, between-physician relative variability is markedly greater for diastolic than for systolic pressures.

Another method of gauging the variability of measurement is by using data collected by replicate measurements of nonsample persons, either by our observers or by other observers. To a limited extent we have done this as part of our regular training. This was also done in a special methodological study which involved (among other things) complete cardiovascular examinations to be replicated on a series of persons.[10] Since

Figure A. Blood Pressures Obtained by Round I Physicians: Deviations From Expected Levels*



*The expected level for a physician is obtained by using the age-sex distribution of his examinees and the average age-sex specific blood pressure levels. Comparisons based on first blood pressures taken.

**Fourth phase.

NOTE: Scales for systolic and diastolic deviations are proportional to their respective standard deviations. The number of persons examined by a physician ranged from 56 to 248, averaging about 120.

neither the Health Examination physicians nor examinees were involved in the methodological study, the results provide only general guidance. They did, however, lead us to discount palpation of the peripheral arteries as a diagnostic finding.

For characteristics which are measured outside the clinic, replicate measurements are feasible and in some cases have been undertaken. Blood glucose and serum cholesterol concentrations are determined at a central laboratory, and for a sample of cases, aliquots are shipped to another laboratory for determination. Some blind replicates have also been introduced into the regular series of laboratory determinations. And, of course, the laboratories also have their own quality controls.

Also evaluated outside the clinic are the electrocardiograms, the chest X-rays and X-rays of the hands and feet. Each of these are evaluated in replicate determinations. The X-rays of the hands and feet may be used to illustrate the problem.

As part of the examination for arthritis and rheumatism, X-rays are taken of the hands and feet of each examinee. These are sent to three physicians especially qualified to evaluate such films. The pair of films for each person is examined independently by each of these three physicians, who note on a standard form any abnormalities observed. In particular, evidence of osteoarthritis is graded from 0 (absent) to 4 (present and severe). Should any two observers assign grades to a film which differ by more than one step, the films for that person are reread in a review session, first independently and then jointly, and a final grade assignment is agreed to. A grade of 2 or more is considered definite osteoarthritis. This is the only evidence used in the Health Examination Survey for a diagnosis of osteoarthritis.

Bypassing problems of etiology, associated symptoms, and medical sequelae (as we do), there remains the grading of the X-ray findings. This is accomplished by first providing a "normal standard" which is X-rayed along with the hands and feet. The standard is a metacarpal bone from an apparently normal person in his 30's. It is encased in lucite and X-rayed along with the hands and feet of each examinee.

Abnormalities once recognized must be graded by comparing the extent of abnormality with a series of graded pictures. For the hands, these pictures are actual X-ray photographs. For the feet, the standard pictures are in the mind of the beholder.

We show data from some of the early readings of Health Examination Survey films (table 4). It will be noted that, despite selection of readers of recognized competency, similar background, and having identical instructions, there is still a distinct difference between the three observers. Not only is the distribution of findings different among readers, but there is a distinct variation between first and second readings by the same

Table 4. Grade Assignments by Different Readers in Replicate Reading of 154 X-rays of the Hands and Feet (Stand 04), Health Examination Survey

Osteoarthritis Grade	First Reading			Second Reading		
	Reader A	Reader B	Reader C	Reader A	Reader B	Reader C
	Number of films graded for hands					
0	13	28	34	16	33	42
1	58	78	67	63	81	48
2	76	41	46	67	33	52
3	5	5	5	6	5	10
4	2	2	2	2	2	2
Average Grade	1.51	1.19	1.18	1.45	1.10	1.23
	Number of films graded for feet					
0	45	55	71	40	48	82
1	50	67	52	49	69	41
2	57	29	29	61	35	29
3	2	3	2	4	2	2
4	-	-	-	-	-	-
Average Grade	1.10	0.87	0.75	1.19	0.94	0.68

NOTE: Readers differ from each other (1% level), but agree with themselves (5% level) in reading pattern.

reader. The within-reader variation is not statistically significant, however, under a hypothesis that distributions of first and second readings are samples from the same universe.

Indeed it may be inquired whether any two interpreters of any medical document ever, in the strictest sense, have the same standards. Thus, if one is to construct a statistical model designed to estimate the prevalence of osteoarthritis, he must first ask what possible meaning he can attribute to this parameter. Surely he can never purport to mimic reality unless he assumes that the parameter itself constitutes a variable. But it is conceivable that if our ambitions are somewhat less we can, in some fashion, define the parameter as an intersection of the idealizations of various observers. We could, for instance, accept as prevalence the expected value of a specified measurement procedure.

Another thing to note about the replicate determinations is that the probability of "correctly" identifying a film varies from film to film. The same series of films was re-evaluated by accident. On the average, the larger the number of observers calling a film abnormal the first time, the larger the number calling it abnormal the second time. (Table 5) If a statistical model is used which presumes a uniform probability of correctly identifying a film, the facts will introduce a correlation between the determinations of the independent observers and spite the model.

Without attempting to suggest at this time a solution to these specific problems, a more general point may be worth making. Statisticians are carefully trained to determine if the data actually satisfy the assumptions of any model they apply to it. The fact is that in a large variety of cases the assumptions are not met, despite the fact that within the specific data this failure may not be demonstrable. If we attempt to mimic the actual complexity of the facts we generally arrive at a point where our model breaks down. A good model should allow us to arrive at two estimates--the first, an estimate (preferably unbiased) of the number of "abnormal" persons; the second, an estimate of the probability that a given person is "truly" abnormal. Such a model should work in a world where different observers

Table 5. Distribution of 154 X-rays of Hands and Feet (Stand 04), by Number of Readers Declaring Film Positive, for Initial Readings of Three Readers, and for Second Readings of Same Readers

Number of Readers Declaring Film Abnormal* in First Reading	Number of X-rays				
	Total	Number of Readers Declaring Film Abnormal* in Second Reading			
		0	1	2	3
Hands					
All cases---------	154	66	29	27	32
0----------	64	57	5	2	-
1----------	34	8	16	9	1
2----------	18	1	6	7	4
3----------	38	-	2	9	27
Feet					
All cases---------	154	84	30	17	23
0----------	90	72	16	2	-
1----------	30	12	14	4	-
2----------	10	-	-	5	5
3----------	24	-	-	6	18

*Evidence of Osteoarthritis of grade 2 or more.

have different definitions of "abnormal" and where the probability of "correctly" characterizing a person varies from person to person. We are still looking for such a model.

6. Estimation

We turn now to the conversion of measurements into estimates of population parameters. The paper does not present a final estimation technique for the HES. It describes a problem, identifies what we consider leading features of the estimation process, and outlines a pattern of thinking and a method of approach to solution of the problem.

The sample design is described in considerable detail in reference 4. It is a multistage stratified probability sample of loose clusters of persons in rather small land segments. Overall it will include about 6,600 persons in approximately the same number of households in some 2,100 land segments in 42 primary sampling units (PSU's) in continental United States. A PSU is a Standard Metropolitan Statistical Area, a county, or a group of 2 or 3 contiguous counties. The master design is self-weighting with respect to persons, but for a number of reasons appropriate inflation factors are not precisely constant for all examinees.

Estimation and calculation of variances must, of course, take into consideration the complex sample design. But for most of the present discussion it will be useful to treat that basic design as though it were a simple random sample.

Consider a simple random sample of n from N individuals in a population for which u_i is the unknown true conceptual value of a characteristic of the ith person (e.g., $u_i = 1$ if the person has hypertension and = 0 otherwise). The population mean value is the parameter $\bar{u}$. An operational approximation to this concept is defined so that if it could be attained it would reflect a value v_i for the ith individual (say, $v_i = 1$ if the individual has prevailing systolic blood pressure ≥ 140 and = 0 otherwise), and $\bar{v}$ for the mean of the population. A survey conducted under a specified set of procedures yields a value x_i ($x_i = 1$ if the measurement obtained for systolic pressure ≥ 140; = 0 otherwise) for the ith unit, and thus is a second order approximation to the conceptual u_i.

The statistician undertakes to process the x-data in such a fashion that they yield estimates of $\bar{v}$ and consequently shed light on judgments concerning the parameter $\bar{u}$. The usefulness of survey data will depend heavily upon the degree of relevance of the v-values to the u-values. In the HES, attention has been given this matter in the selection of u-values which are measurable and of v-values which medical experts agree are indicators of the corresponding u-values, and which can in practice be standardized. Preparation for decisions in this area included contract pilot studies of methods and processes, extensive medical consultations, and two full-dress field trial rehearsals of tentatively chosen content of examination.

The relevance of v-values to u-values is necessary, but only pays off when the x-values from field measurement are reasonably faithful representations of the defined v-values. We have just referred to some of our efforts to deal with this measurement problem.

Sampling and Stratification. In what may be termed the classical approach to estimation, it is assumed that membership of the universe is completely known; sampling is without flaw; response is perfect; the x-measurements are unique, and are taken, recorded and processed without error. The population mean $\bar{x}$ is estimated by $\bar{x}'$, a linear combination of x-observations.

Algebraically this process is summarized with the equation

$$\bar{x}' = \frac{x'}{N} = \frac{1}{N}\left[\frac{N}{n}\sum_{i=1}^{n} x_i\right] \qquad (1)$$

in which the sample observations have been weighted or adjusted by the reciprocal of the sampling fraction. Its sampling variance is well known and, under a broad range of conditions, $\bar{x}'$ is normally distributed about $\bar{x}$.

Many alternatives to this "classical" estimator are known. No attempt is made here to encompass the total field of reasonable possibilities. Note is taken that a worthy objective is to seek an estimator which has not more than a modest bias and which has a relatively low mean square error. Toward this objective, the effect is being explored of application to HES sample observations of other adjustment factors in addition to the sampling weight. Three "types" of adjustment are considered, although all three have much in common.

Ratio Estimation. First, for an item such as hypertension, there clearly are differences in prevalence among different age-sex groups. Since there are available from the Census good independent estimates of population by age and sex, there is a gain in applying a second adjustment

for age and sex control through ratio estimation. This might yield the estimator $\bar{x}''$, where

$$\bar{x}'' = \frac{x''}{N} = \frac{\sum_a \sum_i \frac{N_a}{\sum_i \frac{N}{n} y_{ai}} \frac{N}{n} x_{ai}}{N}, \qquad (2)$$

in which: x_{ai} is observation for ith person in ath age-sex class.

y_{ai} is unity for all persons in ath age-sex class, and equal to zero otherwise, and

N_a is control number of persons in ath age-sex class.

More generally, if $x'_{a\lambda}$ is the basic inflation estimate for any $\lambda^{\underline{th}}$ subclass of the $a^{\underline{th}}$ group, and y'_a the sample inflation estimate of population in the ath class, then the ratio estimate of rate for the total $\lambda^{\underline{th}}$ subclass becomes:

$$\bar{x}''_{\cdot\lambda} = \frac{x''_{\cdot\lambda}}{y''_{\cdot\lambda}} = \frac{\sum_a \frac{N_a}{y'_a} x'_{a\lambda}}{\sum_a \frac{N_a}{y'_a} y'_{a\lambda}}. \qquad (3)$$

The estimate of the aggregate $x''_{\cdot\lambda}$ has lower variance than would the corresponding inflation estimate given sufficient correlation between $x'_{a\lambda}$ and y'_a. Note that this ratio estimation has a type of effect similar to that which would arise from stratification by age and sex, with proportional allocation. For substantially sized sample groups, the ratio estimate is effectively unbiased.

The form in equation (3) has the operational advantage that consistent estimates of $x''_{\cdot\lambda}$ can be secured for any subclass λ simply by adding sample data $x_{a\lambda i}$ which have been weighted by the fixed multiplier $(\frac{N_a}{y'_a} \frac{N}{n})$.

Other Auxiliary Data. Another type of possible adjustment of sample observations is suggested by the fact that several additional items of demographic and health data collected for each of the 6,600 examinees through household interview, are also collected in another NHS survey for a much larger sample of 390,000 persons. This circumstance makes it possible to utilize the larger survey for weighting the smaller in essentially the same way that double-sampling or poststratification is sometimes employed, but at a nearly zero additional cost. This process is called another type of adjustment, although in one frame of reference it is not different from the age-sex ratio adjustment just considered. The order of relative reduction in variance through introduction of such a process is given by the statistic:

$$R = \frac{(H - 1) B}{nW + (H - 1) B} \qquad (4)$$

where

H is the number of pseudostrata,

W is average within-stratum variance, and

$B = \frac{1}{H - 1} \sum n_h (\bar{x}_h - \bar{x})^2 - W$ is between-stratum variance on a unit basis, in which $\bar{x}_h$ is the mean of the $h^{\underline{th}}$ of H pseudostrata, n_h is number of persons in the $h^{\underline{th}}$ stratum, and all calculations are from sample data.

Thus the procedure will be helpful to the degree that (H - 1) B is large as compared with nW.

The use of the 390,000 sample to weight the 6,600 sample could introduce a bias into the process if the two surveys used different procedures. Since the two surveys were carried out under practically identical instructions, with the same auspices, and in large part with the same interviewers, we believe the risk of bias from this source is trivial.

Nonresponse. Earlier in this paper we have discussed the importance of nonresponse in the HES survey. At this point, we stress a hard fact: there is no way in which danger from nonresponse can be entirely eliminated. What one does is to adopt a course which he judges is a reasonable compromise among bias, variance, cost, and operating feasibility. In the HES, extensive steps were taken to minimize nonresponse, and substitute measurements for nonrespondents have been explored as possible techniques. But in essence, the estimation procedure will impute to nonrespondents the data for respondents. The question then resolves into one of what subclasses shall be recognized in the imputation procedure. We suggest that three guidelines are useful in resolving this question: (1) The classes should be ones in which variation in key statistics between classes is large compared with variation within classes; (2) the response rates between classes should be different; (3) the number of respondents in each class should be large enough to avoid letting any respondent represent too

Table 6. Listing of Headings and Stubs for Exploratory Tabulations From Round I Data of the Health Examination Survey[a]

Tabulation Number	r - class	c - class	Number of rc - cells	Avg. value of n_{rc}
1----------	Age-sex (14)	Health Class[b] (4)	56	40
2----------	Age-sex (8)	PSU (14)	112	20
3----------	Age-sex (14)	Income Class (3)	42	50
4----------	Age-sex (14)	Population Density (5)	70	30
5----------	PSU (14)	Income Class (3)	42	50
6----------	Population Density (5)	Geographic Region (3)	11[c]	200
7----------	Super-stratum (11)[c d]	Race (2)	22	100
8----------	PSU (14)	Persuasion Utilized[e] (2)	28	80

[a]Numbers in parentheses indicate number of categories in the classification.

[b]Presence of cardiovascular and arthritic conditions according to interview.

[c]Some cells are vacant.

[d]Super-stratum is a combination of population density and region.

[e]Degree of effort invoked to induce response.

many nonrespondents. Fortunately these considerations are similar to those governing the choice of pseudostratification, and the nature of the appropriate adjustment is similar.

Round I - Experimentation. The 42 HES stands consist of 3 Rounds of 14 stands. Each round is a probability sample of the U. S. Data from Round I are being tabulated separately, partly to produce a few preliminary survey results, but especially to study variations in data for each of eight key statistics by a variety of cross classifications. The cross-tabulations are those indicated in Table 6. For each row, column, and cell, response rates and prevalence rates of each of the eight statistics are being calculated and will be analyzed for relevance to the problems which we have just discussed.

7. Closing Comment.

Our remarks necessarily have concentrated on but a few of the statistical features of the HES. The account, however, has identified a considerable range of problems. We hope to have conveyed the impression that in the National Health Survey we are giving real attention to some of them. We hope also to stimulate others to seek solutions.

REFERENCES

1. U. S. National Health Survey. *Origin and Program of the U. S. National Health Survey.* Health Statistics, Series A-1. PHS Publication No. 584-A1. Public Health Service. Washington, D. C., May 1958.

2. U. S. National Health Survey. *The Statistical Design of the Health Household-Interview Survey.* Health Statistics, Series A-2. PHS Publication No. 584-A2. Public Health Service. Washington, D. C., July 1958.

3. U. S. National Health Survey. *Concepts and Definitions in the Health Household-Interview Survey.* Health Statistics, Series A-3. PHS Publication No. 584-A3. Public Health Service. Washington, D. C., September 1958.

4. U. S. National Health Survey. *Plan and Initial Program of the Health Examination Survey.* Health Statistics, Series A-4. PHS Publication No. 584-A4. Public Health Service. Washington, D. C., May 1962.

5. Commission on Chronic Illness: *Chronic Illness in a Large City: The Baltimore Study.* (Chronic Illness in the United States Vol. IV) Harvard University Press 1957.

6. Commission on Chronic Illness: *Chronic Illness in a Rural Area: The Hunterdon Study.* (Chronic Illness in the United States Vol. III) Harvard University Press 1959.

7. Chen, E. and Cobb, S.: *Further Study of the Non-Participation Problems in a Morbidity Survey Involving Clinical Examination.* Journal Chronic Diseases 7:321-331. April 1958.

8. U. S. National Health Survey. *Attitudes Toward Cooperation in a Health Examination Survey.* Health Statistics, Series D-6. PHS Publication No. 584-D6. Public Health Service, Washington, D. C., July 1961.

9. U. S. National Health Survey. *Cooperation in Health Examination Surveys.* Health Statistics, Series D-2. PHS Publication No. 584-D2. Public Health Service, Washington, D. C., June 1960.

10. U. S. National Health Survey. *Evaluation of a Single-visit Cardiovascular Examination.* Health Statistics, Series D-7. PHS Publication No. 584-D7. Public Health Service, Washington, D.C., December 1961.

MEASUREMENT OF WATER QUALITY THROUGH A NATIONAL SAMPLING NETWORK

Gordon E. McCallum, U.S. Public Health Service

I am pleased to participate in this, the 122nd annual meeting of your American Statistical Association. I am sure that the need for statistical data on which to base man's progressive social, economic, health and welfare requirements is as old as civilization itself. Nevertheless, it came as no small surprise to me that this organization has existed since the year 1840. I know that you and your predecessors -- in your workings with industry, with government, and with organizations and institutions of many kinds -- have contributed much to the growth and development of America. In all probability, basic data such as only the professional statistician can provide, will continue to be in short supply as long as mankind is still inquisitive about the Earth and the Universe in which he lives.

Today, as you see on your programs, I am to speak to you on "The Measurement of Water Quality Through a National Sampling Network."

Background and Present Status of the Water Resource

Water resources development in the United States, encompassing all sectors of society, is truly in the "big business" category, presently involving an expenditure of about $10 billion a year. The federal budget for this fiscal year for water resources investigations and research alone totals $79.6 million, a 24 percent increase over 1962. Fifteen government agencies or departments are involved.

The Federal program which I represent is that of water supply and pollution control, administered by the Public Health Service and its parent agency, the Department of Health, Education, and Welfare.

And now, let us review briefly the background and present status of the water resource itself.

Since the turn of the century, while our population has a little more than doubled, there has been an eight-fold increase in the use of water. Our population, now in excess of 185 million, is using water at the rate of 325 billion gallons a day. Water use is expected to double by 1980 and triple by the year 2000. It goes without saying that as water is used, more liquid wastes are created. There is now nearly six times as much waste -- pollution -- in our rivers, lakes and streams as 60 years ago.

The wise management of our water resource has assumed a position of number one importance. Prominently involved is the abatement, control and prevention of pollution. While sanitary engineers have long recognized and repeatedly voiced the need to protect the quality of our Nation's waters, the public-at-large has been slow to heed the warning and accept the responsibility. The urgency of the need is quite evident when we measure the potential available water supply against projected water requirements for municipal, industrial, agricultural, and recreational use. (These predictions are shown in Figure 1).

Here the complexity of the situation becomes immediately obvious. First, it is clear that reuse of water is unavoidable. Even now, for example, the waters of the Ohio River are used 3.7 times during periods of low flow before they join the Mississippi. Similarly, in the Mississippi River just below St. Paul, more than 10 percent of the flow has recently been through the sewers of Minneapolis or St. Paul one day in 20.

Secondly, criteria for water quality must be developed on a basis which is equitable to all water users. President Kennedy has succinctly stated our national goal as one "to have sufficient water sufficiently clean in the right place at the right time to serve the range of human and industrial needs."

Virtually every use of water, whether for cooling an industrial process or carrying away land drainage, etc., impairs its quality to some extent. This in turn reduces its value to the next user. The expanding need for water, therefore, emphasizes the necessity of preserving its quality as demand rapidly approaches a final limit of supply.

More and more we are finding that water pollution is affected by new factors. In earlier years we were dealing largely with problems of pollution resulting from man's body wastes and from sewage-like industrial wastes. These, and their effects on water, have for the most part been well defined. We have known for years, for example, that waters receiving domestic wastes undergo organic enrichment which results in the production of large quantities of algae, eliminates fish, and destroys the usefulness of the water for recreation, domestic supply, and other purposes.

Today the streams are being invaded by many new and unfamiliar types of wastes to which we might apply the term "persistent," products largely of the new technologics of the past 20 years. Among those with which we are most concerned are synthetic organic chemicals (such as household detergents and agricultural pesticides) and radioactive materials.

These newer wastes, together with residual materials remaining after treatment of conventional wastes, are raising questions as to just what is happening to the quality of the Nation's waters. These questions are coming from water

works operators, agriculturists, recreational groups, conservationists and public officials. The growing importance of water supplies to meet municipal, industrial and agricultural demands, makes it necessary that we know and are able to predict the quality of water much more precisely than we do today.

Defining water quality becomes a most difficult concept. How clean is clean to water's multiple users? And at what point in its role of waste carrier does a river's water become "dirty" to each user? There are many water users, and to each, water quality may have a different meaning. Moreover, the natural mineral and organic content of waters differs from one part of the country to another. Whatever the variations in water quality, its management will be increasingly important to obtaining maximum use from our available water resources.

To carry out such management, basic data on water quality is essential. The Congress recognized this need in the Federal Water Pollution Control Act of 1956 when it gave to the Public Health Service the responsibility to "collect and disseminate basic data on chemical, physical and biological water quality and other information insofar as such data or other information relate to water pollution and the prevention and control thereof." This, the legislation states, is to be done in cooperation with other Federal, State, and local agencies.

The National Water Quality Network

The National Water Quality Network was established in 1957 as a part of the over-all basic data program of the Public Health Service in response to the charge of the Congress. The Network has grown from 51 stations in its first year of operation to 125 stations located on major waterways of the country, with plans for eventual expansion to about 300 stations. Participants include more than 100 local water, sewage or other public utilities, health departments, industries, and universities, State water pollution control agencies, and resident engineers of federal reservoirs. Active local participation is important in this operation. It assures maximum development of all information valuable both locally and nationally. The State and local agencies perform most of the conventional chemical analyses and collect water samples for the more complex examinations. The Public Health Service, in turn, performs the more complex determinations and makes the results available to the participants. In addition, the consultation, training facilities, and other resources of the Public Health Service are available to the cooperating agencies.

The basic data program as a whole, is designed to assemble, examine, and interpret the facts which enable water pollution control agencies and others concerned to determine the scope and character of problems to be solved. It is this last function of the Public Health Service basic data program which I would now like to discuss with respect to water quality.

Objectives

The objectives of the National Water Quality Network are:

1) To maintain continuous intelligence on the nature and extent of pollution affecting water quality.

2) To determine trends in water quality as affected by: (a) water pollution control activities; (b) water resource development activities; and (c) water use and reuse.

3) To provide data on water quality useful in the development of comprehensive water resources programs.

4) To provide data which will guide State, interstate and other agencies in their water pollution control programs, and in the selection of sites for legitimate water uses.

5) To provide data of likely importance to epidemiological and toxicological studies.

Scope of Analytical Activities

Only after careful screening of needs in water resource development was the pattern set for analyses of water samples. Of interest presently in the water quality picture are more than 15 physical and chemical parameters, including radioactivity, plankton populations, coliform organisms and a dozen or more biochemical, chemical, and physical measurements, such as color turbidity, temperature, alkalinity, hardness, dissolved oxygen, etc. Certain of these -- radioactivity, coliform organisms, etc., for example -- are analyzed weekly, others monthly and semi-monthly. Determinations are made also about twice during each year for trace elements in composited samples.

Organic chemicals are adsorbed on activated carbon from about 5,000 gallons sampled over about a week's period each month. In the laboratory the organic materials are extracted from the carbon and separated into about 10 fractions, with specific identifications made where possible.

A rapidly developing area of water quality analysis is that of biology. Plankton have been counted and identified, along with protozoa and certain bacteria, including coliform organisms. Expansion into the determination of fish populations and benthic organisms is under way.

This systematized, continuing collecting of basic data is thus providing the most complete picture we have ever had of the quality of our surface waters.

Programs are also under way to provide de-

pendable electronic measuring and recording equipment to monitor several parameters pertaining to water quality at certain stations. This program will provide a more complete picture of the water at those stations than is now available, but will also provide vast quantities of data which must be reduced to usable form.

This brings us to the important role you, the statisticians, play -- the processing of data to make it usable and understandable.

Data Handling

With but five years of data available from our as yet limited number of stations, we can make only limited determinations of trends and correlations as yet. But, of course, we must be working to sort and solve the problems which must be overcome to enable complete analyses of figures.

Defining water quality is one major task confronting us. What is the quality of our Nation's waters? Is it improving or degrading? Simple questions, but the answers, unfortunately, are far from simple, since no two water users have the same concept of water quality (Figure 2). Domestic users, for example, are concerned with the bacterial purity of their water supply; steel manufacturers are not. High pressure boiler use cannot tolerate a high degree of hardness in water; photographic processors, on the other hand, are concerned not at all with this aspect of the water chemistry. Turbidity is acceptable in low pressure boiler feed water, but not so in most textile manufacturing. Sulfates in water are of little importance to the domestic user, but are certainly unwanted by the pulp and paper manufacturer and must be considered by the farmer irrigating his crops.

Thus, the defining of water quality depends very much on the different water uses. Also, any analysis of water quality must include the natural elements present in waters, and these may vary greatly geographically and geologically. (Figure 3). New Englanders, for example, accustomed to crystal clear recreational waters, may look askance at midwesterners swimming in waters green with algal bloom. Economics may enter the picture, as in certain areas where hard waters affect the use of soap. It has been determined that the savings in the cost of soap will pay for softening of water above 150 ppm hardness.

Selection of sampling location. Care must be taken to locate the Water Quality Network sampling stations where they can continuously provide representative samples of the stream water, and where dependable local assistance can be enlisted in collecting and analyzing samples for the desired parameters. Sampling locations must also satisfy one or more of three criteria: (a) Major waterways used for public water supply, propagation of fish and wildlife, recreational purposes, and agricultural, industrial, and other legitimate uses; (b) Interstate, coastal, and international boundary waters; (c) Waters on which activities of the Federal Government may have an impact.

A site which might answer one requisite, may fail in another. For example, in the lower Mississippi River from Memphis to New Orleans, there are no locations where proper local assistance can be had at sites providing ideal sampling conditions otherwise. Flood is a major problem here, and in those locations where this problem has been overcome, questions exist as to the effects of tributary streams on the collected sample.

The maintenance of usable data presents questions. In what quantities should specific data be collected? How significant are they in the over-all assembly? The chemist, the biologist, and the engineer jointly must decide the importance of present analyses, when to abandon old or begin new analyses. But only the statistical analyst himself can cope with problems involving too-sparse data or the accumulations of large quantities of unprocessed data. The latter may be expected to grow rapidly as automatic instrumentation is installed to monitor water quality.

Constant evaluation of laboratory techniques is important to the successful operation of the National Water Quality Network. Two questions arise in this connection: (a) What is the most dependable technique of analysis for a given parameter, and (b) is the performance of all laboratories such that the data should be published for use by other organizations?

A basis for these evaluations is provided by the Analytical Reference Service of the Public Health Service's Training Program at the Robert A. Taft Sanitary Engineering Center in Cincinnati. It prepares samples for analysis by various interested laboratories throughout the United States, including those participating in the Network. The results are evaluated for laboratory performance and dependability of methods on the basis of the known contents of the samples. Table I shows one such evaluation involving the occurrence of several metals in water samples. Other direct evaluations of techniques are carried out by the Water Quality Section and by the Research Branch of the Division of Water Supply and Pollution Control.

The statistics obtained from analyses made by participating laboratories are presented graphically (Figure 4). These include the standard deviation from the amount of substance added and a 50 percent range. Each laboratory receives a summary report of results. If a participating laboratory falls short of the acceptable limit of accuracy set by our chemists, the laboratory is alerted and corrective suggestions offered to it.

The utilization of water quality data is, of course, the ultimate aim of the Public Health Service's sampling Network. Analyses of data for trends, cycles, unusual values, ecological

correlations, and epidemiology directed toward a better understanding of our national water resource is our specific task.

A computer program is being developed to analyze trends, cycles and unusual values. Variations in data from one sample to the next and from year to year may make difficult the detection of such trends and cycles, although some data do lend themselves to such analysis. Figure 5, for example, shows the cyclic nature of plankton population at two stations on the Yellowstone and Columbia Rivers. It also shows upward trends in plankton populations at two stations on the Illinois and Mississippi Rivers. Such elegant and well-defined data are not available at all Network stations, however.

Our data may be interpreted on levels of: (1) the single station, (2) the single main stem of a river, (3) the basinwide system, or (4) the nationwide picture.

Whereas the Figure 5 data demonstrate the cyclic nature of plankton populations found at some stations, and trends found at others, in Figure 6 we see a striking example of a trend in the chloride data collected at Yuma, Arizona.

Plankton data similar to that of Figure 5, but for two stations on the Potomac River, are shown in Figure 7. The plankton populations at the upstream station, Williamsport, are lower and less variable than at the Great Falls Station, which demonstrates a high degree of variability and frequent high populations.

A different type of data presentation is made in Figure 8 for the green alga, Scenedesmus, in the Missouri River.

Populations of coliform bacteria, for years accepted as indicators of fecal contamination, are presented in Figure 9 for the Missouri and Mississippi Rivers. It may be noted that the Missouri enters the Mississippi at St. Louis.

Radioactivity analysis best illustrates data interpretation for the Nation as a whole. Depicted in Figure 10 are the quarterly medians of 47 National Water Quality Network stations operated since the beginning of the Network. Statistics have been applied to test significant changes in radioactivity at some stations, but have not as yet been applied to the data of Figure 10.

Conclusion

It is obvious to a group such as yours that the engineering profession -- and the sanitary engineers in particular -- are faced with a severe problem of defining water quality, of understanding the factors influencing this quality, and of predicting future needs. This appraisal is a great responsibility. The very health of the Nation -- both physical and economic -- is at stake. Our methods of measurement of water quality must be increased in number, improved in sensitivity, and expertly translated into terminology which will give the public adequate information on how clean or how soiled our streams and lakes may be.

The statistician with an understanding of biology and chemistry can definitely contribute to the solution of our many problems. It is urgent that improved approaches to data-handling as well as interpretation be developed continuously in order that the needs of health agencies, industrial groups, conservation interests, and ultimately, the individual citizen can be satisfied.

References

(1) Analytical Reference Service, Public Health Service, U.S. Department of Health, Education, and Welfare. "Analytical Reference Service - Water Minerals No. 3" (1961)

(2) Crow, James F. "Mechanisms and Trends in Human Evolution" Daedalus 90:416-431 (1961)

(3) Kroner, R. C., Ballinger, D. G., and Kramer, H. P. "Evaluation of Laboratory Methods for Analysis of Heavy Metals in Water" J. A. W. W. A. 25:117-124 (1960)

(4) U.S. Senate, Select Committee on National Water Resources "Water Quality Management" in Water Resources Activities in the United States, Print No. 24, (February, 1960) p. 9

(5) U.S. Senate, Select Committee on National Water Resources "Water Supply and Demand" in Water Resources Activities in the United States, Print No. 32, (August, 1960)

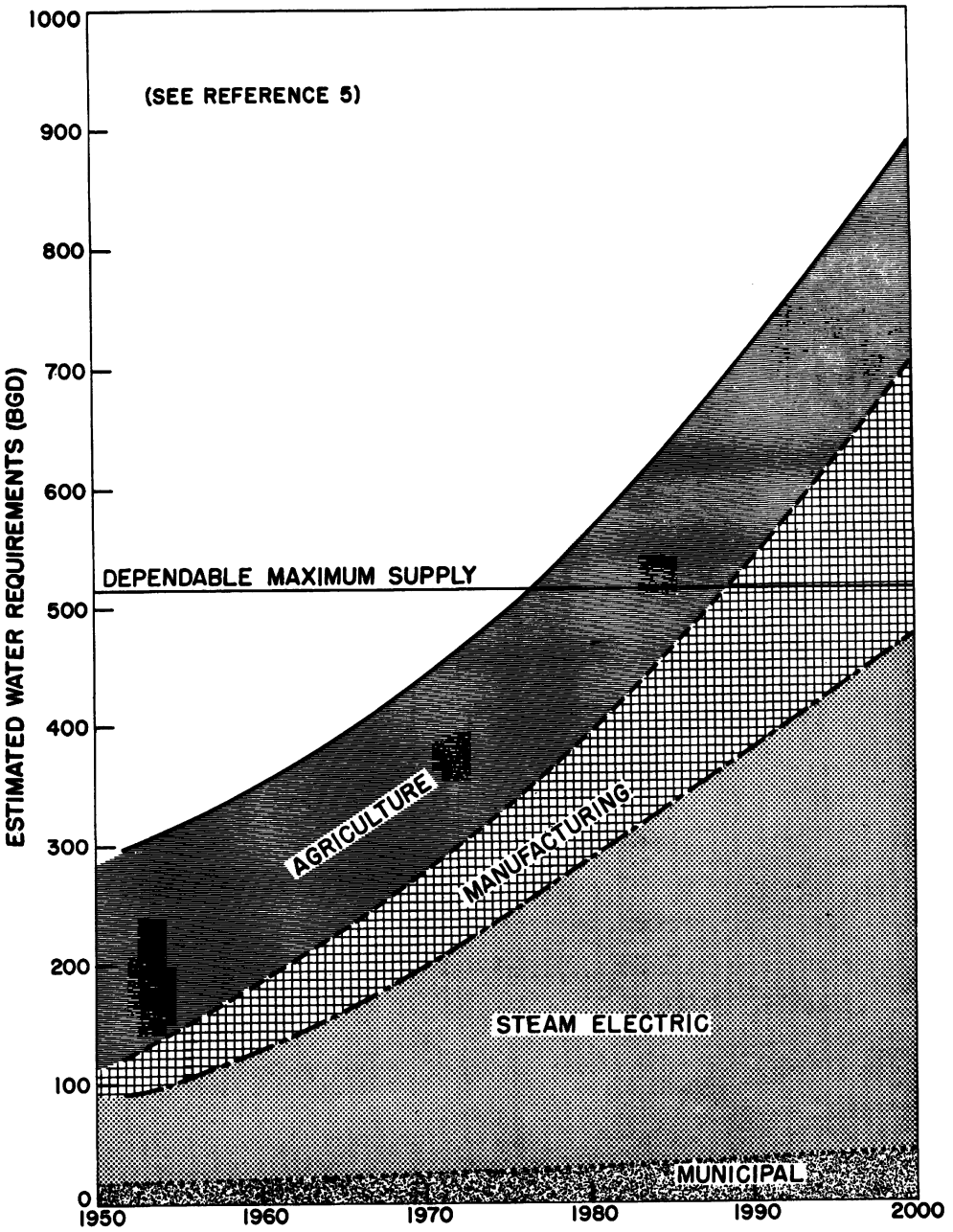
(6) Water Quality Criteria. State Water Pollution Control Board, Sacramento (1952)

TABLE I

(After Kroner et al, Reference 3)

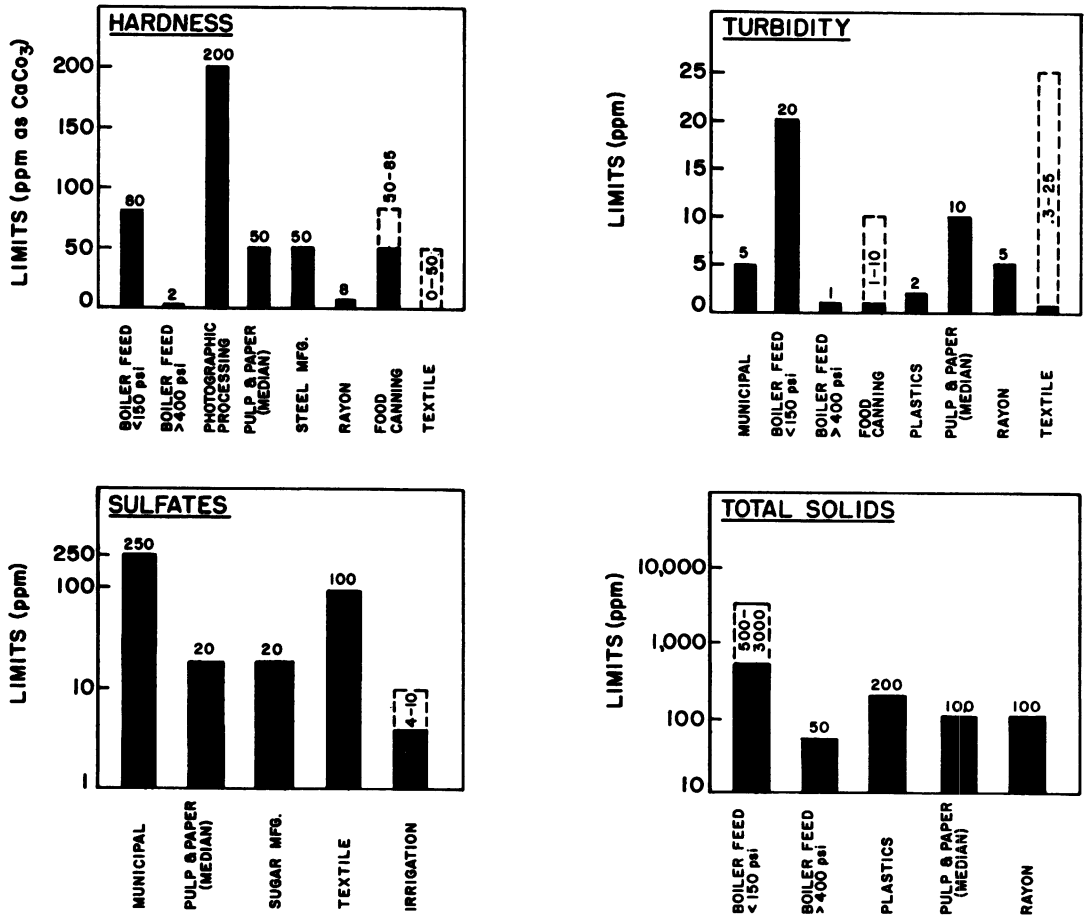
Determination	Method	No. of Laboratories Reporting	Amount Added mg/l	Avg. of Amounts Observed mg/l	Standard Deviation From Amount Added	
					mg/l	percent
Iron	all methods	17	0.55	0.55	0.126	22.3
	o-phenanthroline	9		0.55		
	bypyridyl	4		0.57		
	tripyridyl	1		0.50		
	thiocyanate	3		0.57		
Copper	all methods	14	0.56	0.60	0.204	36.4
	carbamate	8		0.58		
	cuprethol	5		0.62		
	dithizone	1		0.56		
Manganese	all methods	17	0.20	0.20	0.100	50.0
	periodate	8		0.17		
	persulfate	8		0.23		
	other	1		0.23		
Aluminum	all methods	12	2.00	3.44	1.880	94.0
	aluminon	9		3.47		
	hematoxylin	1		2.52		
	other	2		3.75		
Cadmium	all methods	7	0.16	0.18*	0.061*	38.1
	dithizone	6		0.17		
	polarograph	1		0.25		
Lead	all methods	14	0.12	0.13*	0.40*	33.3
	dithizone	13		0.13		
	polarograph	1		0.18		
Chromium	diphenylcarbazide**	14	0.15	0.13	0.041	27.3
Zinc	dithizone**	11	7.47	7.73	0.979*	13.1

* One atypical result not used in calculation
** Only method used



PROJECTED WATER NEEDS FOR MAJOR WATER USERS IN THE UNITED STATES

FIGURE 1



VARIATIONS IN WATER QUALITY NEEDS OF INDUSTRY AND MUNICIPALITY

FIGURE 2

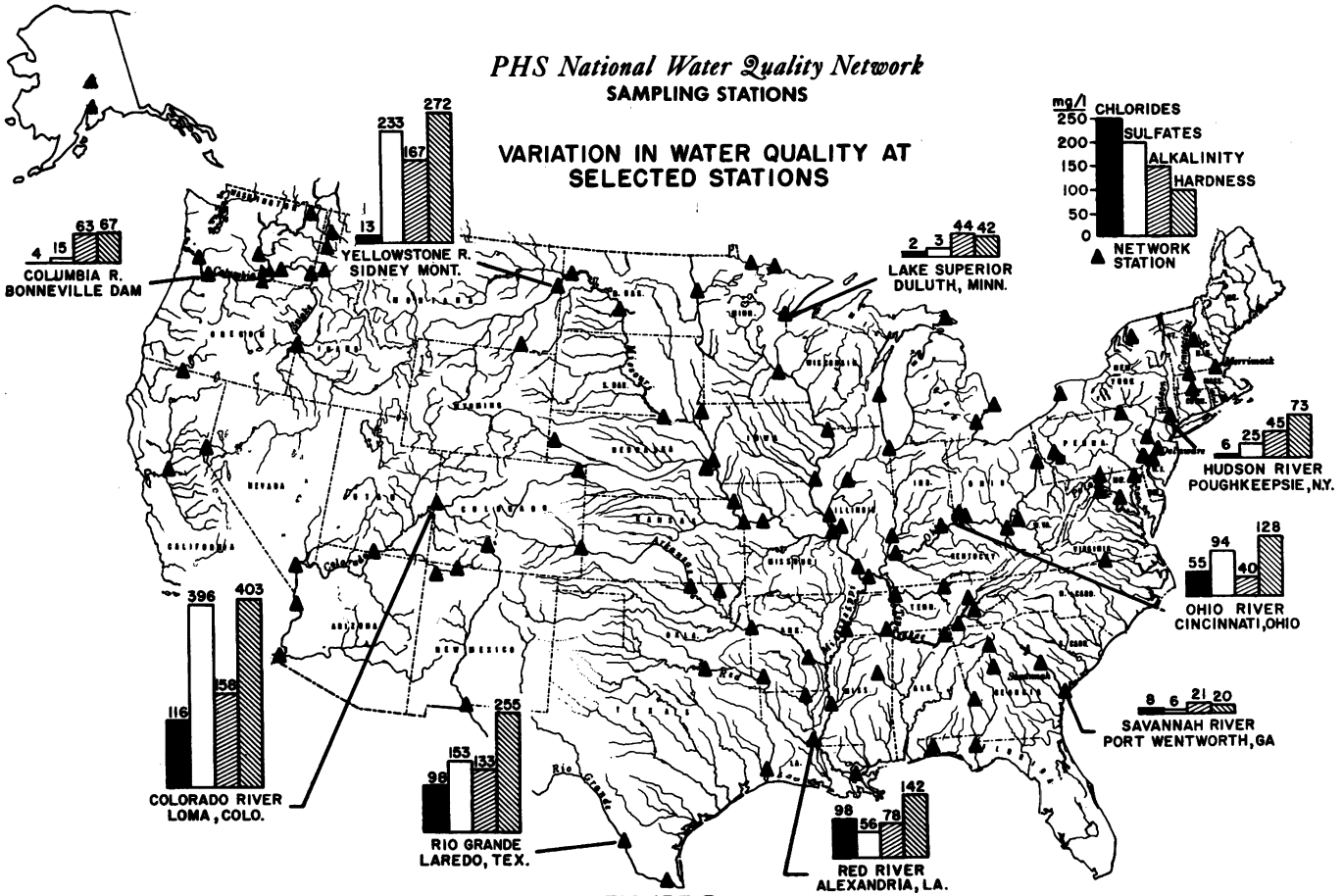
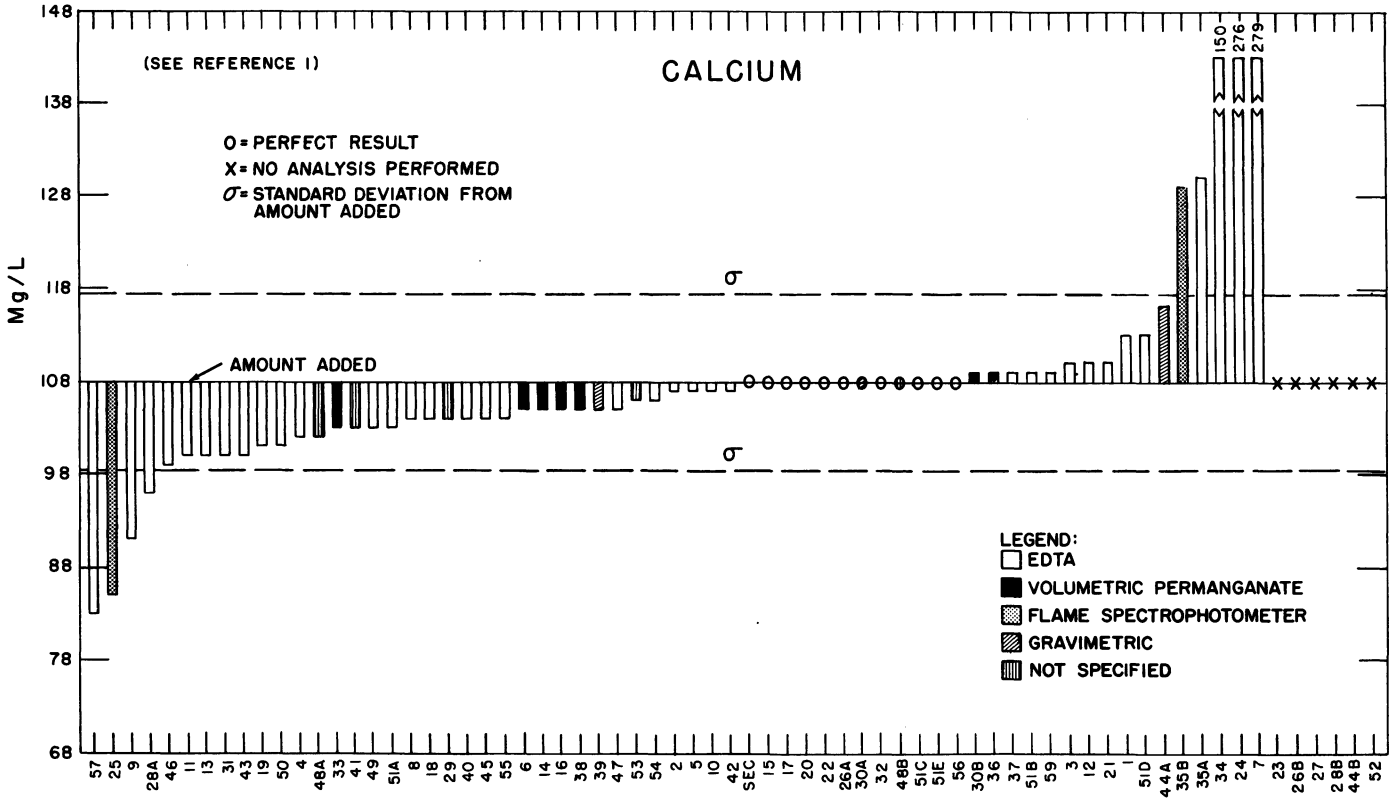
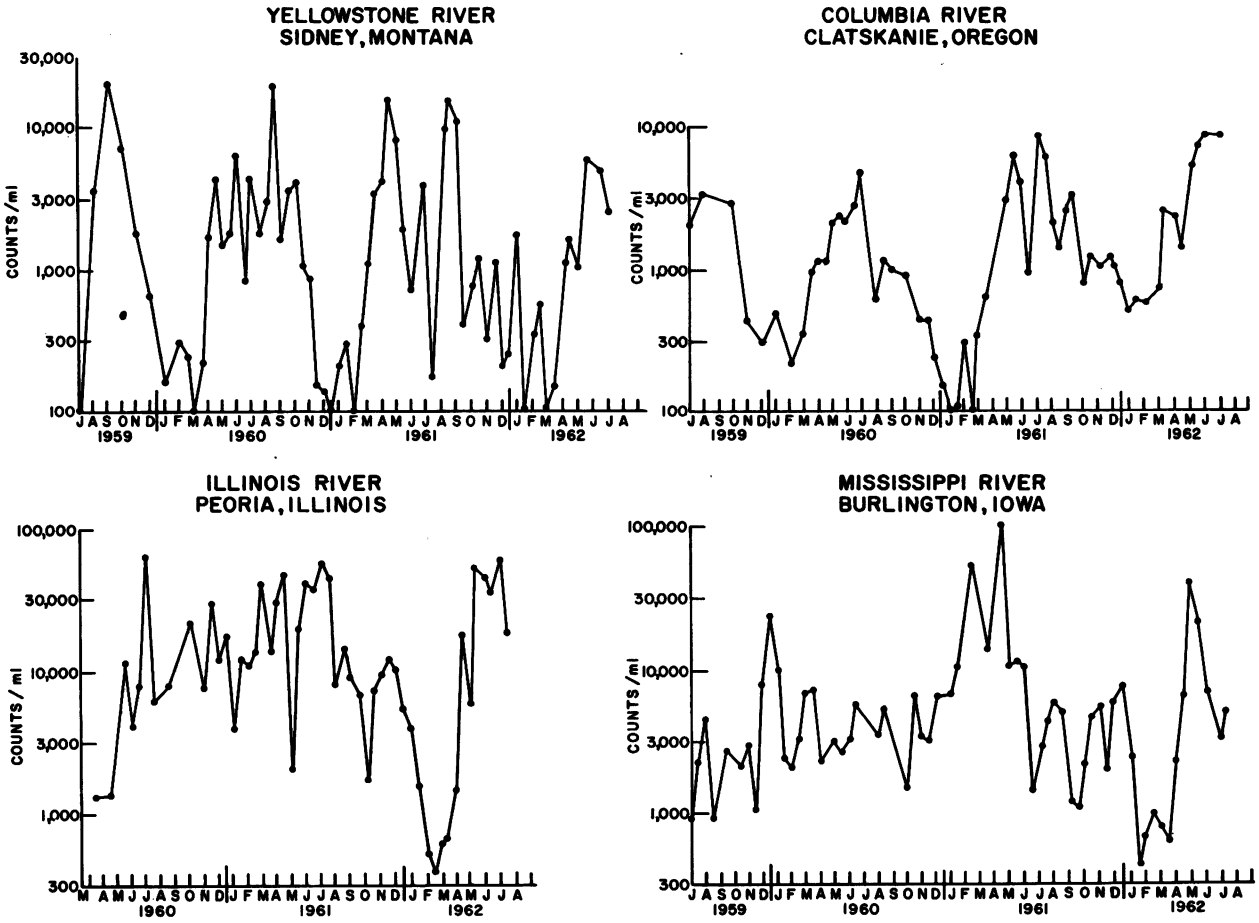

PHS National Water Quality Network
SAMPLING STATIONS
VARIATION IN WATER QUALITY AT SELECTED STATIONS
mg/l
250
200
150
100
50
0
CHLORIDES
SULFATES
ALKALINITY
HARDNESS
NETWORK STATION
63 67
4 15
COLUMBIA R.
BONNEVILLE DAM
272
233
167
13
YELLOWSTONE R.
SIDNEY MONT.
44 42
2 3
LAKE SUPERIOR
DULUTH, MINN.
73
45
25
6
HUDSON RIVER
POUGHKEEPSIE, N.Y.
128
94
55
40
OHIO RIVER
CINCINNATI, OHIO
8 6 21 20
SAVANNAH RIVER
PORT WENTWORTH, GA
396
403
116
156
COLORADO RIVER
LOMA, COLO.
255
153
139
98
RIO GRANDE
LAREDO, TEX.
142
98
56
78
RED RIVER
ALEXANDRIA, LA.

FIGURE 3

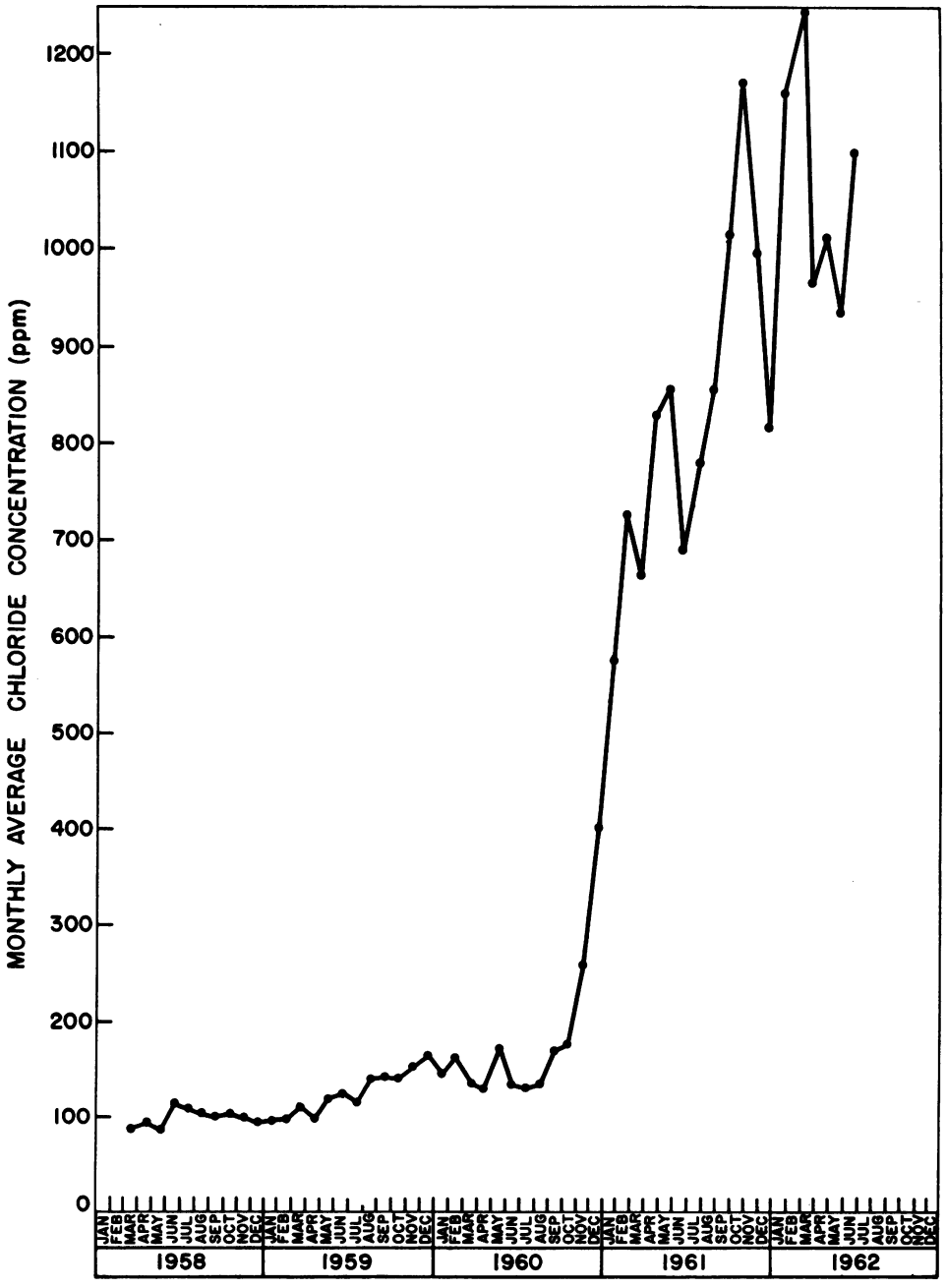


ANALYTICAL REFERENCE SERVICE

FIGURE 4

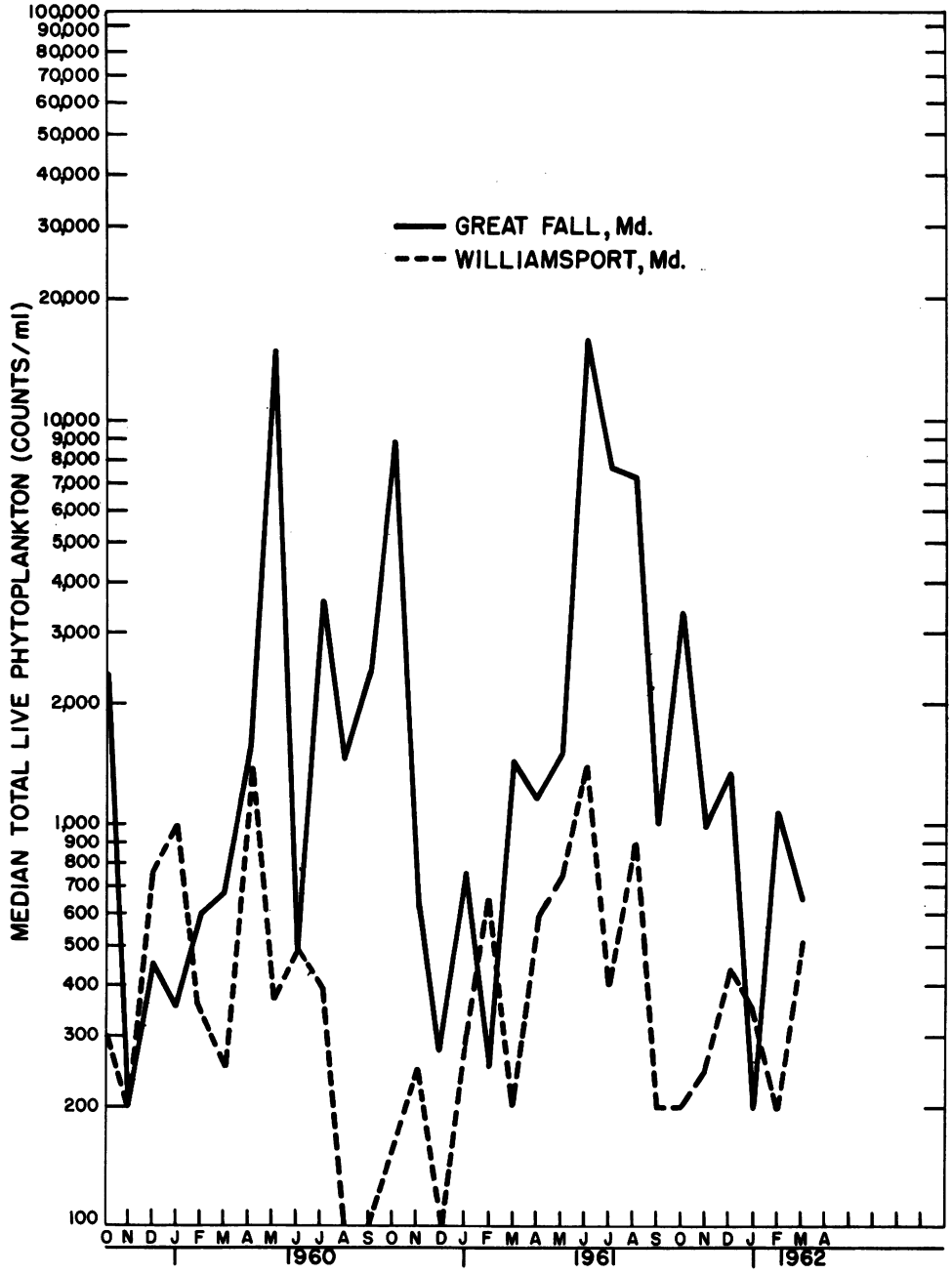


CYCLES AND TRENDS
TOTAL LIVE PHYTOPLANKTON COUNTS
FIGURE 5



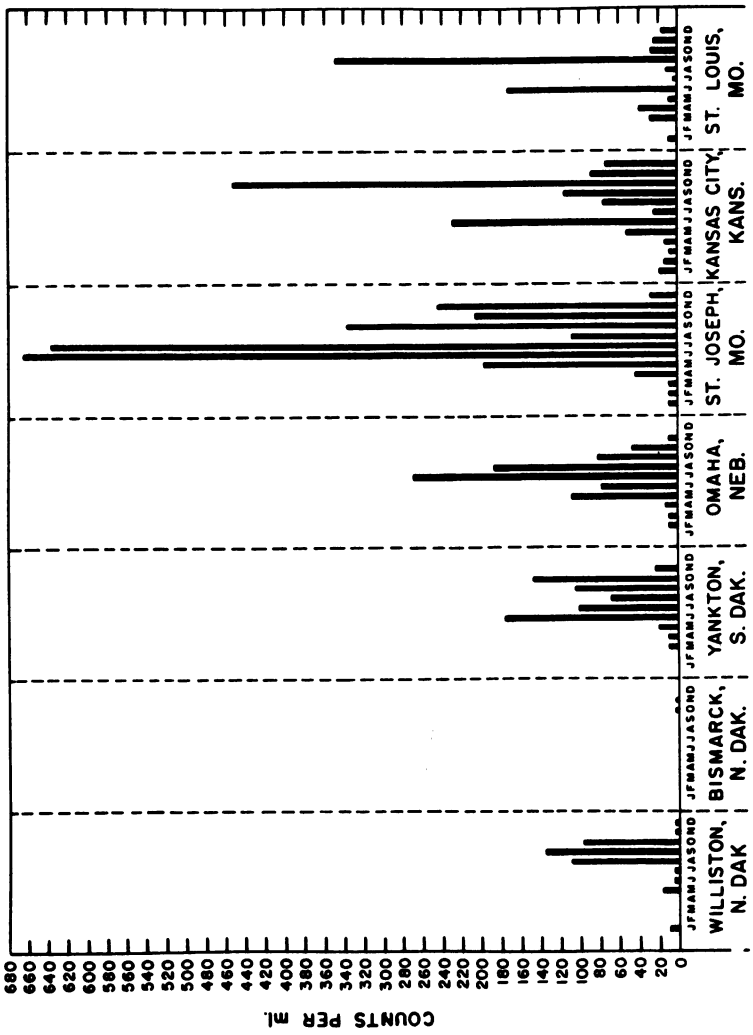
COLORADO RIVER AT YUMA, ARIZONA
CHLORIDE VS TIME

FIGURE 6



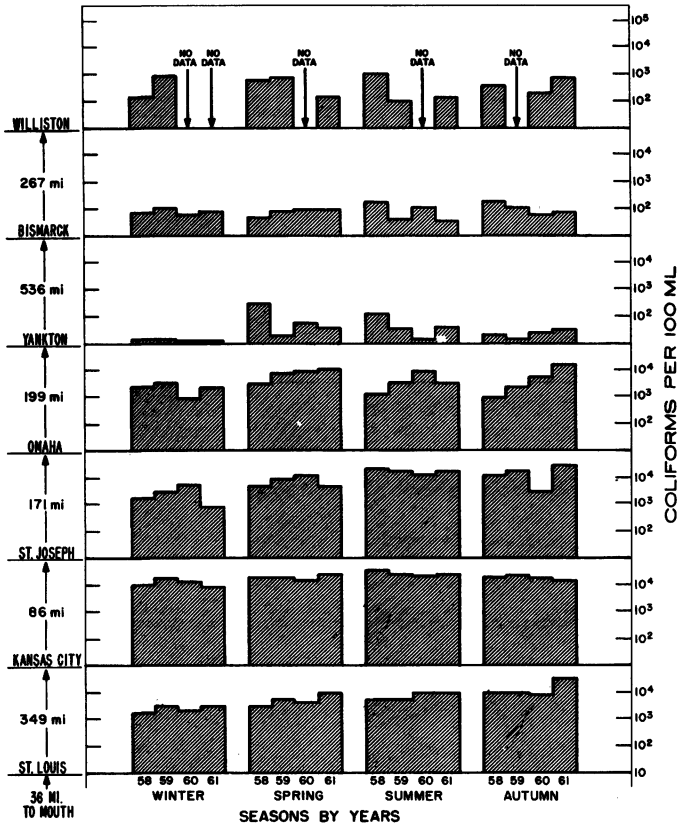
POTOMAC RIVER
MEDIAN MONTHLY TOTAL LIVE PHYTOPLANKTON

FIGURE 7

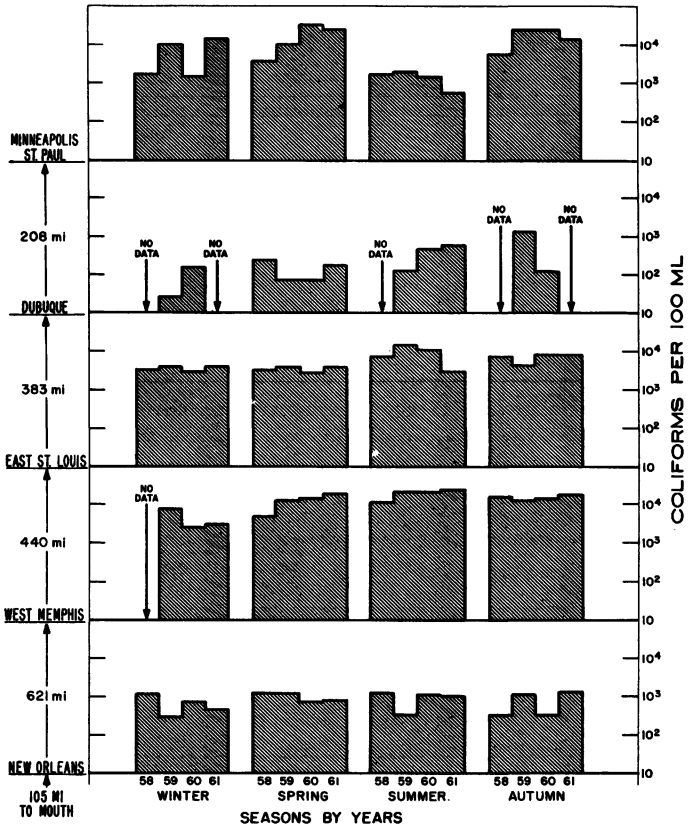


SCENEDESMUS
AVERAGES FOR 1958, 1959 AND HALF OF 1960
MISSOURI RIVER

FIGURE 8

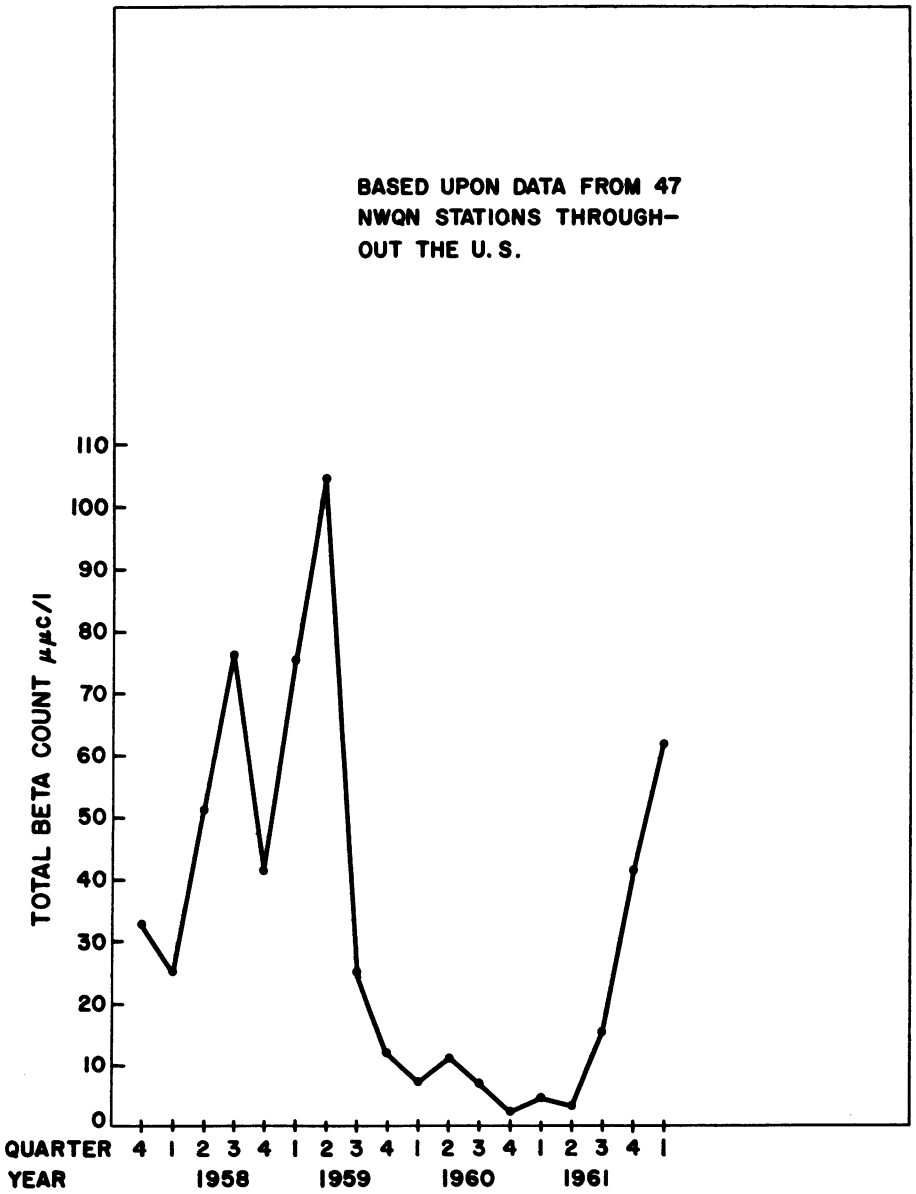


SEASONAL COLIFORM MEDIANS
MISSOURI 1958-1961



SEASONAL COLIFORM MEDIANS
MISSISSIPPI 1958-1961

FIGURE 9



MEDIAN QUARTERLY TOTAL BETA RADIOACTIVITY IN THE U.S.

FIGURE 10

VI

TECHNIQUES OF DATA COLLECTION

Chairman, Jacob J. Feldman, National Opinion Research Center

Page

Comparison of Three Information-Gathering Strategies in a Population Study of Sociomedical Variables - Joseph R. Hochstim, California Department of Public Health..... 154

Interviews Hardest-to-Obtain in an Urban Health Survey - Regina Loewenstein, John Colombotos and Jack Elinson, Columbia University..... 160

On the Accuracy of Recording of Consumer Panels - Seymour Sudman, National Opinion Research Center..... 167

Discussion - Reuben Cohen, Opinion Research Corporation..... 176

COMPARISON OF THREE INFORMATION-GATHERING STRATEGIES IN A POPULATION STUDY OF SOCIOMEDICAL VARIABLES

Joseph R. Hochstim
California State Department of Public Health

The Human Population Laboratory is an epidemiologic study whose primary interest is in learning the distribution of disease in a community and examining the relationships between disease and those social, psychological and environmental factors which we loosely call "way-of-life".

Since the development of disease, particularly chronic disease, is a long-term process, a longitudinal study design with its repeated interviews of the same individuals over a period of time appears to be appropriate for the Human Population Laboratory.

Obviously, it is costly to conduct personal interviews with a reasonably large sample of the population repeatedly, particularly if one follows all or a sample of the migrants wherever they go. This consideration led us to look into less costly methods of information gathering, primarily telephone interviews and mail questionnaires, or to some combination of methods which might lower costs without reducing quality.

Our first step was to comb the literature to discover what studies had been made on the comparative merits of personal interviews, telephone interviews and mail questionnaires. We went through the major American publications in survey research, applied psychology, sociology, statistics and marketing. In many of the journals we made a complete search back to 1948.

In general, much has been written about the various methods of data collection, but attempts to compare these different methods have been very few. The limited information available does not help in developing a longitudinal study in a compact geographic area, where rigorous controls can be exercised over the quality of methods chosen.

We therefore found it necessary to conduct our own study. This report represents the results of this study. Our objective was to compare three strategies of information gathering - mail, telephone and personal interviewing used in certain combinations - in terms of rate of return, the completeness of the returns and cost. In addition, we aimed to investigate comparability of the results obtained by the three strategies.

The site selected for the Human Population Laboratory is Alameda County, in the San Francisco Bay Area. The county has a population of almost a million, mainly urban and suburban, like most of California. The population is heterogeneous with respect to occupation, socioeconomic class, race and other factors possibly related to disease occurrence. The area is small enough to allow data collection with a minimum of travel expense, while its proximity to the State Department of Public Health permits a maximum of supervision

A short description of the sample design is necessary here because this design affected the analysis of the study.

An area probability sample of Alameda County, California, was drawn based on 1960 census data. A multistage design was used. The only 1960 census statistics available at the time of sampling were census enumeration districts; so we decided to use these as our primary sampling units. The county was stratified by geography and the enumeration districts were selected with probability proportionate to size. Each enumeration district in the sample was subdivided into secondary units - blocks or, in rural areas, quasi blocks. Two blocks were chosen from each sample enumeration district, again with probability proportionate to size. Within each block, a cluster of six households was drawn from a random start. Altogether, 175 enumeration districts, including 350 blocks and about 2,100 households, were thus selected. This sample, supplemented by a sample of new construction, conversions and demolitions undertaken since the 1960 Census, yielded a total of 2,148 housing units, representing about 1 out of every 150 housing units in Alameda County.

Table 1

RATE OF RETURN OF HOUSEHOLD ENUMERATION

SAMPLE STATUS OF HOUSING UNITS	HOUSING UNITS	
	Number	Percent
Total Housing Units	2,148	100
Enumeration Completed	1,973	92
Enumeration Not Completed	175	8
Reason not completed		
Refusal	33	2
Not at home	20	1
Other biasing reasons[1]	10	a
Vacant	90	4
Other nonbiasing reasons[2]	22	1
Total Effective Sample (Excluding "Vacant" and "Other nonbiasing")	2,036	
Number and Percent of Effective Sample Completed	1,973	97

[1] Includes households containing persons who were too ill to be interviewed or who were senile.

[2] Includes persons ruled out of the sample by definition e.g., students living on campus and military personnel on military bases.

[a] Less than 1 percent.

The 2,148 sample households were then enumerated. Enumeration involved a listing of all household members by sex, age and relation to head, as well as certain housing and other information. As you see in Table 1, the enumeration was successfully completed in 92 percent of the selected households; or in 97 percent of the occupied housing units.

Altogether 1,973 housing units were enumerated in groups of 6 housing units per block. These housing units were next subdivided into two samples. All those units which fell on the even-numbered lines of the block listing sheets became Sample A, the subject of our present discussion. The remaining households were reserved for a replication of the study design, but using a different subject matter.

The sample for the present study was divided into three subsamples, each a representative sample of the total county. Out of the total of 350 blocks, 50 were selected systematically for personal interviews, 3 households per block. In the other 300 blocks, 3 households per block also were taken - one assigned at random to the telephone strategy, the other two assigned to the mail strategy. Obviously, cost factors determined the disproportionate numbers assigned to the three strategies.

For all three strategies we did everything possible to achieve a high rate of return. We used all the prestige of the Department and the fact that we are promoting public health. Each household received an advance notice of its selection for the study, in the form of a letter from the Chief of the Human Population Laboratory, personalized as far as possible and hand-signed. Extreme care was exercised at every stage of the study to induce a sense of participation in the respondent.

In all three strategies every member of the household seventeen years and over was eligible for the study. Identical questions were used throughout, the topics being demographic, familial, behavioral and medical.

Let me describe each of the three strategies. In the mail strategy, each eligible member of the household was sent a separate questionnaire, with an accompanying letter. A second mailing, again with an accompanying letter, was sent to those who did not return the first questionnaire; and, if necessary, a third - this by certified mail with the request that the addressee sign a return receipt. The postoffice also was asked to report a new address in case the person had moved.

Those still not responding were then called upon, either by telephone or in person. Callbacks were made to obtain as high a return rate as possible within the limits of the budget. So much for the mail strategy.

In the second strategy, the primary aim was to conduct the interview by telephone. However, some people do not have a telephone. In order to keep the three samples comparable, the telephone sample included its proper share of nontelephone subscribers; otherwise it would have represented a higher socioeconomic group. Those households drawn for the telephone strategy which did not have a telephone were handled like the mail sample.

The third strategy employed personal interviews, with callbacks where necessary.

The overall rate of return from all three strategies combined was ninety percent (Table 2). Nonreturns were made up of seven percent refusals and three percent not-at-home and unable to locate.

Table 2

RATE OF RETURN OF DATA COLLECTION

SAMPLE STATUS OF PERSONS	PERSONS	
	Number	Percent
Total Adults[1] in Housing Units	1,984	100
Interviews or Questionnaires Obtained	1,779	90
Not Obtained	205	10
Reasons not obtained		
Refused	135	7
Not at home	15	1
Unable to locate (moved and left no forwarding address)	33	1
Other reasons	22	1

[1] Persons 17 years old and over

Looking next at the 3 strategies separately (Table 3), we find that a completed questionnaire or interview was obtained from 88 percent of the mail strategy sample; 91 percent of the telephone strategy sample was completed, as was 93 percent of the personal interview strategy.

Table 3

RATE OF RETURN BY STRATEGY AND METHOD OF COMPLETION

METHOD OF COMPLETION	STRATEGY		
	Mail	Telephone	Personal
Number of Adults in Households	1,109	571	304
Number of Interviews Obtained	977	518	284
	Percent		
Percent Interviews Obtained	88	91	93
Obtained by			
Original method	81	72	90
Other methods	7	19	3
Mail	x	14	3
Telephone	4	x	-
Personal	3	5	x

x Not applicable.

In each of the three strategies the great bulk of the assignment was completed by the method originally selected. For example, the 88 percent obtained in the mail strategy was made up of 81 percent returned by mail, the other 7 percent divided about evenly into telephone and personal interviews. The 91 percent return from the telephone strategy included 72 percent completed by telephone, 14 percent by mail and 5 percent in person.

Our next concern was the completeness of the questionnaires. Even though the mail strategy yielded a satisfactory rate of return, the questionnaires themselves might have a disproportionately high rate of unanswered questions. Over the telephone as well as in the personal interview, the interviewer can encourage responses; this sort of prodding is not possible in the mail questionnaire. Therefore, it was expected that the proportion of unanswered questions in the mail strategy might be higher than in the other two strategies. As it turned out, this expectation was justified, but the nonresponse to individual questions was so low - seldom going over five percent even in the mail strategy - that it did not present a problem.

Table 4

SUMMARY OF NONRESPONSE TO SIXTY-ONE QUESTIONNAIRE ITEMS BY STRATEGY

PERCENT "NO ANSWER"	NUMBER OF ITEMS NOT ANSWERED		
	Strategy		
	Mail	Telephone	Personal
0.0	20	25	30
0.1-0.9	14	13	14
1.0-1.9	10	14	5
2.0-2.9	4	3	4
3.0-3.9	2	3	2
4.0-4.9	5	3	4
5.0 or More	6	-	2
Average Percent "No Answer"	1.9	0.9	1.0

I should comment here that the figures on rate of completeness for the mail strategy include a few edited questionnaires. If a questionnaire was returned which had, say, a whole page left blank, it was assigned to an interviewer who telephoned the respondent to obtain the answers to the unanswered questions. Thus, the rate of completion of mail questionnaires is the rate of respondents' answers plus some subsequent prodding in the few cases where there were gross omissions.

Coming now to the question of costs: A great many expenditures enter into the conduct of a survey - administration, planning, sampling, questionnaire construction, testing, etc. We are concerned here only with the cost of interviewing, on the assumption that most other expenses will not vary greatly with the strategy employed. However, there are questions of just what charges to include in interviewing, and what fractions of certain costs to assign to each strategy.

You will remember that we conducted a household enumeration before doing the actual interviewing, primarily in order to get names and addresses for the telephone and mail strategies. One could take the position that enumeration and interviewing might have been conducted simultaneously in the personal strategy and therefore none of this enumeration cost should be assigned to this strategy. However, this view is not quite realistic because all members of the household 17 years and over were to be interviewed, and even where a respondent was available on the enumeration trip, additional visits were often necessary to interview the other eligible household members. In fact, in many cases the enumerator found no one at home and simply took name and address for subsequent telephone calls to complete the enumeration. Consequently, it is not unreasonable to charge at least a fraction of the enumeration to the personal strategy.

Another expense involves selection, training and supervision of interviewers. Such costs are quite substantial in the personal interview strategy, less so in telephone strategy and minor in the mail strategy. How much of this interview supervision should be charged to each strategy?

We computed per interview costs on a number of bases, ranging from charging personal strategy with none of the enumeration to charging various fractions, and from dividing interviewer training and supervision costs equally among the three strategies to assigning them heavily to personal and telephone strategies. Depending on what assumptions we made on these issues, we came up with the following comparisons: Telephone strategy interviews cost from ten percent to twenty percent more than the mail strategy returns and personal strategy interviews cost from two to two-and-a-half times as much as mail.

I want to draw a broad band of caution around these comparisons. These figures apply only to this particular study, and even here we are not sure how good our cost accounting was. Also, with a different rate paid to interviewers, or with a different length of questionnaire, or with any difference in procedure - for example, no certified letter in the mail strategy - cost ratios would change, perhaps considerably. All we can say is that, not unexpectedly, the personal interview strategy cost substantially more than either of the other two.

So far we have shown that in terms of rate of return and in terms of completeness, the three strategies were quite comparable and that in terms of cost, the mail and telephone strategies were more efficient than the personal interview strategy.

The next item to investigate is a comparison of the findings from the three strategies. Let us look first at those data for which comparisons with outside sources are available, i.e., Census. Does any one of the three strategies appear to have a marked advantage over the other?

Not all statisticians agree on the usefulness of comparisons between sample data and population data. While some are unimpressed, others think that such comparisons are a realistic way of assessing the representativeness of a sample. Without taking sides on this issue, I will report the comparisons we made on a series of items obtained in the Human Population Laboratory which were also reported in the 1960 Census.

When the returns from the individual strategies are compared with one another and with census data, we find (Table 5):

1. In most cases results from the three strategies are in good agreement with each other and with census, and
2. None of the three strategies is consistently closest to the census on all items compared.

Table 5

COMPARISON BETWEEN U.S. CENSUS 1960 AND HUMAN POPULATION LABORATORY STUDY 1961

DEMOGRAPHIC VARIABLE	U.S. CENSUS 1960[a]	HUMAN POPULATION LABORATORY STUDY 1961 Strategy		
		Mail	Telephone	Personal
Total Persons 17 Years and Over	595,556	977	518	284
	PERCENT			
Sex				
Male	47	48	46	45
Female	53	52	54	55
Age				
Less Than 25	13	10	13	13
25-34	20	19	24	22
35-44	22	23	22	22
45-54	18	19	15	19
55-64	13	13	13	13
65-74	9	9	8	5
75 and Over	5	6	4	5
Not Reported	-	1	1	-
Race				
White	86	86	86	78
Negro	11	10	11	16
Other Races	3	4	3	6
Nativity				
Native Born	91	88	90	90
Foreign Born	9	11	10	10
Not Reported	-	1	b	b

DEMOGRAPHIC VARIABLE	U.S. CENSUS 1960[a]	HUMAN POPULATION LABORATORY STUDY 1961 Strategy		
		Mail	Telephone	Personal
	PERCENT			
Marital Status				
Married	71	75	75	68
Widowed	9	7	8	9
Divorced	5	5	4	4
Separated	2	2	3	4
Never Married	13	11	10	14
Employment Status				
In Civilian Labor Force	61	61	59	63
Employed	57	57	54	58
Looking for work	4	4	5	5
Not in Labor Force	39	39	41	37
Occupation[1]				
Professional and Managerial	23	21	21	21
Clerical and Sales	25	29	30	31
Craftsmen	14	13	16	13
Operatives and Laborers	20	20	19	18
Service Workers	11	10	12	15
Uncodable	7	7	2	2
Industry[1]				
Agriculture	1	2	3	2
Construction	6	6	7	5
Manufacturing	22	21	22	19
Transportation, Utilities, Communication	8	8	9	12
Trades	18	16	16	21
Services	31	26	28	27
Public Administration	8	8	11	10
Industry Not Reported	6	13	4	4

[1] Census figures refer to employed persons, Human Population Laboratory Study excludes housewives and students.

[a] Adjusted to exclude students and military personnel not in the Human Population Laboratory Study.

[b] Less than 1 percent.

Interestingly, the mail and the telephone strategies show just as close relationships to census as does the personal strategy. In fact, the largest difference - though not statistically significant - appeared in the personal strategy. The important point, however, is that all three strategies are close to census and to one another.

Next we proceed to a comparison of the substantive findings obtained by the three strategies. I won't go into the details here of computation of the standard errors except to say that I have been using a technique pretty much like the one Leslie Kish presented in his article on "Confidence Intervals for Clustered Samples", in the April 1957 issue of the American Sociological Review. This technique is not unlike the one Jerome Cornfield described in the study on "Health and Medical Care in New York City". In essence, these techniques take cognizance of the fact that we are dealing with a cluster sample instead of a simple random sample.

Generally speaking, the three strategies drew similar responses. In the great majority of findings, the differences observed are not large enough to be statistically significant. There are, however, a number of questions and a number of items on which statistically significant differences did appear. By chance alone, we would expect that in five percent of the comparisons, a statistically significant difference would occur. In fact, we observed such differences on six percent of the items.

If we look at the questions where significant differences are found, and if we arrange them systematically and classify them in order to seek a rationale for what may have happened, certain patterns emerge.

First, we find that about one-third of all the differences appear in a series of six questions out of sixty-one. This was our first attempt to develop an index of physical activity and I am afraid we were not successful.

Second, some of the interstrategy differences seem to arise from the need for special instructions or explanations. For example, interviewers can be given detailed instruction as to how occupation will be coded and should be recorded. The fact that in the mail strategy seven percent of the answers were uncodable, while only two percent each for the telephone and personal strategies were uncodable, documents this speculation. The same point holds for recording "industry" and for some other items too.

Third, there are situations where ideas about acceptable responses give the advantage to one strategy over another. Herb Hyman and our chairman, here, in their standard work on "Interviewing in Social Research", have dwelt on the importance of interviewer expectations in survey research. Respondents, too, have ideas of what is expected or is acceptable. If somebody asks me, or any one of you, "How do you feel?" we almost automatically answer, "Fine," even if we don't feel very well. Similarly, when our interviewers asked respondents how they would generally rate their health - excellent, good, fair or poor - the automatic response tended to be "excellent". However, asked to put down a check mark on a questionnaire, the response tended to be less off-the-cuff than in the more social situation.

Table 6

RESPONDENTS' HEALTH RATING BY STRATEGY

	STRATEGY		
	Mail	Telephone	Personal
Total Persons	977	518	284
	Percent		
Total Percent	100	100	100
Health Ratings			
Excellent	30	37	44
Good	51	51	45
Fair	17	10	10
Poor	2	2	1

Comparing findings on health rating for the three strategies, you will observe in Table 6 that the proportion saying "excellent" is much higher in the personal strategy, and the proportion saying "fair" is higher in the mail strategy. I would hypothesize that in this case, the mail strategy may come closer to reflecting the respondents' true state. What is operating here is probably not so much a desire to deceive as the impulsive face-to-face response compared with the more considered reply to a printed question.

Another example: In our culture, drinking alcoholic beverages, particularly by women, is not quite acceptable middle-class behavior. If we look at the findings on frequency of drinking wine, beer and hard liquor (Table 7) we notice that some eight percent to thirteen percent more women tell the interviewer that they never drink than make this response on the impersonal mail questionnaire or over the telephone. This appears to be a case where respondents conceivably may answer more honestly if they are not facing an individual who may be critical of their behavior.

Table 7

FREQUENCY OF DRINKING ALCOHOLIC BEVERAGES
WOMEN RESPONDENTS BY STRATEGY

	STRATEGY		
	Mail	Telephone	Personal
Total Women	507	282	157
	Percent		
Total Percent	100	100	100
Women Saying They			
Never drink wine	46	44	55
Never drink beer	51	49	59
Never drink whiskey or liquor	36	34	47

So much for the pattern of differences. Looking now at the magnitude of the differences, we find them not very large. Aside from the physical activity items, the mean of the statistically significant differences between any two strategies is seven percentage points and the range is from four to fourteen points.* Introducing this element of the size of the observed differences brings up the question of statistical in relation to meaningful significance. Often a difference, even though statistically significant, has little substantive importance.

To sum up: We all know that in any interview study, however conducted, some respondents are easier to find than others, and that the hard-to-get respondents are very costly. Because we have been aware that certain types of people do not respond to mail questionnaires or are not available by telephone, we tend to conduct entire studies by personal interview, even when the information may be obtainable by other techniques.

What I have reported here is a study of three strategies of information gathering. In two, we tried to obtain the easy-to-get interviews in the less costly ways, at the same time identifying the hard-to-get respondents for follow-up by the more expensive techniques. The third strategy was based on the personal interview in order to gain comparative data.

As I reported: First - rate of return was satisfactory for all three strategies

RATE OF RETURN BY STRATEGY

	PERCENT
Mail Strategy	88
Telephone Strategy	91
Personal Interview	93

Second - We did get the bulk of our interviews by the method originally assigned

RATE OF RETURN
STRATEGY BY METHOD OF COMPLETION

METHOD OF COMPLETION	STRATEGY		
	Mail	Telephone	Personal
	Percent		
Total	100	100	100
Mail	92	15	3
Telephone	4	80	-
Personal	4	5	97

Third - Rate of questionnaire completeness also was satisfactory for all three strategies

AVERAGE PERCENT "NO ANSWER"
PER QUESTIONNAIRE ITEM

Mail Strategy	1.9
Telephone Strategy	0.9
Personal Interview Strategy	1.0

* On the physical activity items, the corresponding figures are 11 percentage points for the mean and 5 to 19 points for the range.

Fourth - Interviewing costs were about two to two-and-a-half times as high for the personal strategy as for the mail strategy, and about twice as high for the personal strategy as for the telephone strategy

COST RATIOS OF DATA COLLECTION
FOR THE THREE STRATEGIES

(Per Questionnaire Cost of Mail Strategy Equals 100)

STRATEGY	RATIO TO MAIL STRATEGY
Telephone Strategy	110-120
Personal Interview Strategy	190-240

Finally - So far as substantive findings go, the returns from the three strategies were generally quite comparable and it appears that on most items the strategies could have been used interchangeably.

Before I begin to sound as though I am advocating the use of the mail or telephone strategy, I should say that I am aware of certain limiting factors. One I have touched on: A public health department, seeking information which presumably would benefit many people, may have an aura that induced an unusual rate of mail and telephone cooperation. Obviously, a lower return would have changed cost comparisons considerably.

Another limiting factor is that certain kinds of inquiries simply cannot be made by mail or telephone - because they are too complicated, because they involve tests of knowledge, because question sequence is important, because of the sensitivity of the subject and so forth.

What we are searching for, of course, is a way that can be used in certain situations to save costs, particularly where follow-ups in a longitudinal study are involved, and where the problem of respondent mobility looms large.

There is much analysis still to be done on these data. Additional investigations are also being made. You may remember that in discussing the sample design, I mentioned that we put aside half of the households enumerated, those listed on the odd-numbered lines of the enumeration sheets. We have just completed in these households a survey which replicates the methodological study I have discussed, with a different subject matter, but a good many identical questions. Therefore, with a sample using households on the very same blocks, we have a two-pronged effort here: a complete replication on some issues and also new information on a very sensitive subject matter.

Since these date are in the processing stage, no comparisons can be made at this time. It is hoped that by replications of studies of this type, it will be possible to build up a body of knowledge giving us information as to if, and under what conditions and with what safeguards, mail and/or telephone interviews, supplemented by personal follow-ups for the hard-to-get cases, are practicable.

INTERVIEWS HARDEST-TO-OBTAIN IN AN URBAN HEALTH SURVEY

Regina Loewenstein, John Colombotos, Jack Elinson
Columbia University School of Public Health and Administrative Medicine

Introduction

Organizations doing interview surveys are always concerned about the possible bias in their results due to the persons in their samples who were not interviewed. For example, the proportion of persons with a particular socio-economic characteristic may be over or under represented among those interviewed. Statisticians are asked to measure the bias due to non-response. Supervisors of field operations would like to know the types of persons who are hardest-to-interview in order to develop techniques to interview them.

These questions were the concern of the staff of a collaborative research project called the Master Sample Survey being carried out by the Columbia University School of Public Health and Administrative Medicine in the Washington Heights Health District in New York City[1].

The Master Sample was designed in 1959 by the Columbia-Washington Heights Community Mental Health Project so that random samples or groups with specified demographic characteristics could be selected for small studies. By early 1960, the plan was to use this sample for a collaborative research project on a variety of medical and health problems. This work was supported by the Health Research Council of the City of New York, under Contract U-1053.

The five research groups participating in the 1960-1961 Master Sample Survey were: Patterns of Medical Care Study of New York City Department of Health, Columbia-Washington Heights Mental Health Project, National Council on Alcoholism, and Cardiovascular and Neurological study groups of the Columbia University School of Public Health and Administrative Medicine.

During 1960, interview schedules were designed by John Colombotos in collaboration with representatives of each of these research groups to answer questions of interest to one or more of them.

From November 1960 through April 1961, interviews were sought from a two-stage stratified cluster sample with 4,500 dwelling units[2]. By the first cut-off date, interviews had been obtained from 3,300 dwelling units or 77 percent of eligible dwelling units, allowing for sampling losses.

The staff wanted to study the possible bias due to the absence of interviews with about one-fourth or 1,000 of the dwelling units in the sample on the first cut-off date. Persons in 550 of these dwelling units had been contacted and refused to be interviewed. The remaining 450 dwelling units had not been interviewed for reasons other than refusal; namely, no contact had been made or some contact had been made but with no direct refusal.

To study the characteristics of this "hard-to-obtain" group, a sample of one-third or 340 of the dwelling units with no interviews were selected for intensive investigation. Between May 1st and July 15th, members of the supervisory field staff, two medical students[3], and several of the best interviewers exerted intensive effort to obtain interviews in these 340 dwelling units.

By the final cut-off date[4], interviews had been obtained in 210, or about 60 percent of the dwelling units. About one-half of the previous refusals and three-fourths of the previous non-refusals had been interviewed. Of the remaining 130 dwelling units with no interviews, almost 80 percent had refused at least once, and only 20 percent had never been contacted. Thus, persons never interviewed were primarily people who had refused to be interviewed at least once.

The 3,300 early interviews gave information about 3,700 families[5] and 9,200 persons. Adding the interviews obtained in the second round, weighted by three, the tabulations of this survey are based on 4,000 interviews with 4,400 families and 10,750 persons. Data about the 4,400 families will be referred to as marginal distributions.

In Table 1 are the distributions of size of family and age of persons of the early and late cases, with the late cases classified as previous refusals and previous non-refusals according to their status as of the first cut-off date. The distributions of these two characteristics are almost identical for the Washington Heights Health District, the estimated sample sought, the early cases, and the marginal distributions.

There are differences in size of family and age of persons between the early and late cases. In spite of this, the marginal distributions and the distributions of the early cases are almost identical because, even after weighting, the 229 families in late cases are only 16 percent of the 4,400 families represented in the marginal distributions.

NOT INTERVIEWED GROUP

Since the survey had been done within one year of the 1960 United States Census of Population, tabulations were purchased from the Bureau of the Census on a few characteristics about persons in dwelling units with no interviews for whom names were known provided that the names

and addresses were exactly the same at the time of the Census. Tables about 94 matched dwelling units gave data on size of family, sex, race and age, of about 200 persons living in these dwelling units.

It was then possible to estimate the distributions of these 4 characteristics for 98 percent of the dwelling units eligible for interview from the sum of (a) the early cases; (b) the late cases, multiplied by 3; and (c) the non-interviews in the one-third sample, multiplied by 3. The characteristics of the remaining 2 percent, that is, dwelling units for which no names were known plus those whose names were not found at the specified address in the Census records, were assumed to have the same characteristics as non-interviews that had been found.

Table 2 summarizes the results of this procedure. The sample sought was estimated to be 4,800 families with 11,500 persons. At the first cut-off date, 77 percent of the families and 80 percent of the persons were represented in the interviews. After weighting the late interviews by 3; 91 percent of the families and 93 percent of persons in the sample were represented.

By the first cut-off date, interviews had been obtained from 73 percent of families of one and two persons, and 83 percent of families of three or more persons. During the second round, the interviewing rate among larger families remained higher, and interviews obtained by the second cut-off date represented 89 percent of families of one or two persons and 96 percent of families of three or more persons.

There were no differences in the proportions of men and women or of white and nonwhite persons who were interviewed early or late, or who did or did not refuse to be interviewed in the first round.

Table 2 also shows differences in ease of interviewing persons in various age groups. By the first cut-off date, 85 percent of persons under 45 years, and 75 percent of persons 45 and older had been interviewed. During the first round, the refusal rate had been highest among persons 65 years and older. In the second round the refusal rate was highest among persons 45 to 64. After the second cut-off date, 97 percent of persons under 45, and 90 percent of persons 45 and older were represented.

The relationship of age, size of family, and field status have not yet been analyzed.

COMPARISONS OF GROUPS BY FIELD STATUS

Information about non-response bias can be obtained by comparing early and late interviews, and also by comparing previous refusals and previous non-refusals, with regard to various characteristics.

Such comparisons have been done for about 200 variables, and the results of these analyses will be summarized. All differences mentioned are significant on the five percent level. Significantly different percentages usually differ absolutely by 5 to 15 percent.

1. Demographic and Socio-Economic Characteristics.

The distributions of the early and late cases and of previous refusals and non-refusals with regard to size of family and age of persons are shown in Table 1.

Families who had refused to be interviewed in the first round included more two-person families than did the early cases and the previous non-refusals. Among the previous non-refusals were more one-person families, reflecting the difficulties interviewers had to find them at home.

Compared with the early cases, the previous non-refusals had more persons 45 to 64 years of age, and the previous refusals had more persons 65 and older.

There were no differences between early and late cases, or between previous refusals and non-refusals with regard to race, sex, religion, rent, family income, public assistance, occupations of employed persons, countries of origin of heads of families, plans to move, reporting no close relatives, or having a family doctor.

Many of the characteristics that had different distributions among early and late cases may be related to the larger proportions of small families or older persons in the late cases. The late cases had more families with no children under 16, with no persons under 35, and at least one person 65 or older; and more widowed persons. Fewer of the adults in the late cases were high school graduates. Re occupation and income - the employed persons in the late cases were more frequently self-employed, earning at least $6,000 a year, and on the same job for at least 10 years. Fewer adults in the late cases worked full-time. Re previous residences - the late cases had more persons born in the United States outside of New York City, and more families who had been at the same address for at least 15 years.

Some of the characteristics which were different in the previous refusals and previous non-refusals may be related to the differences in family size and age. The previous refusals had more families with all persons over 65 years, more adults who kept house, fewer adults who worked full time, and more adults with no high school education. The health insurance coverage was less comprehensive for the persons in the previous refusals.

Other characteristics which differed among previous refusals and non-refusals may be related to the larger proportion of persons born in West-

ern Europe among previous refusals. While the field work was in progress, the interviewers reported difficulties in obtaining interviews with some refugees from Nazi governments because of their fear to let strangers into their homes. Among persons not born in the United States, the percentage who became citizens was higher among the previous refusals than the previous non-refusals. Also, the previous refusals had more heads of families who had lived in New York City for at least 15 years. More previous refusals reported that their closest relative was other than a parent, child, or sibling.

2. Reported chronic conditions, etc.

Early and late cases, and previous refusals and non-refusals did not differ with regard to the reported presence during the past year of 29 chronic conditions, 12 out of 13 impairments, 4 symptoms of cardiovascular conditions, drinking problems, institutional care, or limitations in ordinary activities due to illness.

These findings seemed surprising because there were more older persons in the late than early cases, and in the previous refusals than the previous non-refusals. Therefore, the proportions of persons in 12 age-sex groups reporting at least one of the 29 chronic conditions were analyzed. In 10 of the 12 groups, the percents of persons in late cases reporting at least one condition were lower than among the early cases, and in 9 of the 12 groups the percents were lower among previous refusals than previous non-refusals.

Were persons in the late cases, especially previous refusals, healthier or more laconic in the interview?

3. Reported medical care during past year.

There were no significant differences between early and late cases in the proportions of persons reporting attended conditions, the mean number of reported visits per person interviewed or per person reporting at least one attended condition.

But a significantly smaller proportion of the previous refusals reported one or more attended conditions. The mean numbers of reported visits to each type of facility were smaller for the previous refusals than the non-refusals, but the differences were not significant.

The late cases and the previous refusals had fewer persons with at least one hospital stay, and more of the stays reported by the late cases were 10 days or more. The reported numbers of days in bed for reported attended conditions were the same for persons in the three groups of cases, but the previous non-refusals reported more persons than did the previous refusals who missed daily activities 20 or more days because of reported attended conditions.

There were fewer persons among the previous refusals reporting 10 or more out-of-hospital visits in one year than among the early cases or among the previous non-refusals.

The proportions of persons reporting that all out-of-hospital care was from doctors in private practice were the same among the early cases, previous refusals, and previous non-refusals. But previous refusals went more often than previous non-refusals to doctors whom they called "family doctors." The previous non-refusals had more visits to out-patient departments and more frequently used three or more different facilities for out-of-hospital care in one year.

Among the previous refusals, in comparison with previous non-refusals, there were more reported attended conditions with the first and only care from a doctor known before, and fewer with any care from a facility referred by a medical person.

Further analyses may help in the interpretation of the differences in reported chronic and attended conditions for persons in the previous refusal and non-refusal cases. How are these differences related to perceived need for medical care, the amount and type of medical care sought and received, and the cooperation of interviewees in the two groups of cases?

4. Questions asked of Sample of Adults.

Questions suggested by the Patterns of Medical Care Study and the Columbia-Washington Heights Community Mental Health Project were asked of a sample of persons 21 years and older in the Patterns Cases.

Marvin Belkin of the staff of the Patterns of Medical Care Study developed 11 indices based on their responses to 37 multiple choice questions. There were no significant differences between early and late cases or between previous refusals and previous non-refusals on seven indices; namely, Awareness of Physician's Social Status, Physician's Interest in Patient's Welfare, Acceptance of Sick Role, Preventive Medical Behavior, Illness-Dependence Orientation, Work Indispensability, and Ethnic Exclusivity.

The late cases had more persons with low indices re Reliance on Friends and also Friendship Solidarity, and previous non-refusals had more with low indices on Friendship Solidarity. On the Index for Family Orientation to Tradition and Authority, the late cases and the previous refusals both had more with high scores.

An index on Attitude toward Medical Care was based on responses to three questions about trying different doctors, doubting doctors can help, and wanting details when ill. These questions were:

"I believe in trying different doctors to find which one I think will give me the best care."

"When I am ill, I demand to know all the

details of what is being done to me."

"I have my doubts about some things doctors say they can do for me."

Persons were asked whether they agreed or disagreed with these statements, and could also give qualified responses.

There were more persons among the previous non-refusals than previous refusals who agreed with all three of these questions, which was classified as lowest on this index. Further analyses may show if a low index on Attitude toward Medical Care is related to more reported medical care.

There were no differences by previous field status as to the reporting of dental care, eye examinations, periodic checkups, polio shots, balanced diet, or expressed need for additional health services. But, the early cases had a higher score on the Knowledge of Illness test than the late cases, possibly related to the differences in education.

Bruce Dohrenwend of the Columbia Washington Heights Mental Health Project suggested the inclusion of 22 questions from the Midtown Manhattan Study[6]. These items were found to discriminate between psychiatric patients and well persons[7]. The items used were originally drawn from the neuropsychiatric screening inventory used during World War II and the Minnesota Multiphasic Personality Inventory. A score was developed from the responses to these 22 questions. There were no differences between persons in early and late cases, or in previous refusals and non-refusals as to the distributions of this score or the average score.

The four field status groups also did not differ in responses to five questions on anger, guilt, and fear.

5. Note re Comparisons.

The previous refusals and previous non-refusals differed from each other on more of the 200 characteristics than did the early and the late cases, because there were many characteristics in which the previous non-refusals were more like early cases than like the previous refusals.

Almost all of the comparisons that have been discussed are first order-relationships, that is, the two pairs of field status groups as a whole have been compared without controlling for other variables. Exceptions were the comparisons of previous refusals and previous non-refusals as to reported chronic conditions by age. Other analyses of second and third order relationships might give further understanding about the types of persons who refused to be interviewed in this health survey, or who were hard to interview for other reasons.

EARLY CASES AND MARGINAL DISTRIBUTIONS

To measure the effect on the final results of the intensive effort of the second round, comparisons with regard to these same 200 variables were made of the distributions of early cases and the marginal distributions; that is, early cases plus three times the late cases.

The distributions of 135 variables differed absolutely by less than one percent in all categories, and 50 variables differed by one to two percent in one category. In the 15 variables with absolute differences of two to three percent, there were usually small bases for the percentages and/or large proportions of no answers to specific questions among the late cases.

That is, the distributions of very few characteristics would have been different if field work had stopped at the first cut-off date, and there had not been intensive effort to obtain interviews from a sample of the non-interviews at that time.

SUMMARY OF FINDINGS

1. By the first cut-off date, interviews had been obtained from 77 percent of 4,800 families; 12 percent had refused to be interviewed and 11 percent had not been interviewed for other reasons.

2. Intensive field effort on a one-third sample of cases not interviewed by the first cut-off date resulted in estimated response rates of 91 percent of families and 93 percent of persons eligible for interviews.

3. Interviews with families of only one and two persons were harder to obtain than interviews with larger families. Interviews with older persons, i.e., 45 years or older, were harder to obtain than interviews with younger persons. Further analyses are needed about the correlation of these findings.

4. Many of the socio-economic characteristics that differed among early and late interviewed cases appear to be related to age or family size. The early and late cases, however, reported the same amount of chronic and attended conditions.

5. One group of persons who were interviewed late had previously refused to be interviewed. This group of "previous refusals" differed from others who were interviewed late in a number of socio-economic characteristics. Some, but not all, of these differential characteristics were related to age or family size. Previous refusals reported fewer chronic and attended conditions than previous non-refusals of the same age.

6. Previous refusals and previous non-refusals did not differ in eight attitudinal areas. Previous refusals had lower indices on Friendship Solidarity, higher on Family Orientation

toward Tradition, and a different Attitude about Medical Care.

7. There were no differences in the distributions of about 200 characteristics between early cases and final marginal distributions, i.e., early cases plus late cases weighted by three. For 135 of these variables the distributions never differed by more than one percent in any category, and there were never differences of more than three percent between the two distributions.

IMPLICATIONS

These findings, of course, would not necessarily be the same in other surveys with different interviewees, interviewers, and topics.

Frequent comparisons of the distributions of important characteristics while a survey is in progress would inform the field staff what groups need more intensive effort. Decisions with regard to additional field work to be done would then depend on the nature and magnitude of the non-response bias, the available funds, and whether the major emphasis in the particular survey is to estimate the characteristics of a community or to study the inter-relationships of these characteristics.

In this survey, estimates of 200 social, health and medical care characteristics for the community as a whole from the first 77 percent of families in the sample were essentially identical with estimates based on interviews representing 91 percent of the families in the sample. Further analyses will show whether inter-relationships of these characteristics would have been effected if interviewing had stopped at the first cut-off date.

FOOTNOTES

1/ According to the United States Censuses of Population and Housing, the Washington Heights Health District had 100,000 housing units and 270,000 persons in 1960.

2/ The sample was designed in 1958 by Marvin Glasser now at Harvard University.

3/ Charles Enzer and Donald Plevy, in cooperation with the 1961 Summer Training Program of the New York City Department of Health.

4/ Cut-off dates are in a sense arbitrary from a research point of view since they are influenced by "deadlines" for reports, time and personnel budgets, and, of course, available financial support.

5/ The number of families is estimated to be higher than the number of dwelling units because more than one family lived in some dwelling units.

6/ Leo Srole, Thomas S. Langner, Stanley T. Michael, Marvin K. Opler, Thomas A. C. Rennie, Mental Health in the Metropolis, New York: McGraw Hill, 1962.

7/ Thomas S. Langner, "Psychophysiological Symptoms and Women's Status in Two Mexican Communities," in Jane M. Murphy and Alexander H. Leighton (Editors), Approaches to Cross Cultural Psychiatry, in preparation.

TABLE 1

Size of Families and Age of Persons of the Washington Heights Health District, of the Estimated Sample, of Interviews received by Field Status, and Estimate of Non-interviews.

	Washington Heights Health District[1]	Estimated Sample	Marginal Distributions[2]	Interviews Received				Not Interviewed[5]
				Early[3]	Late[4]			
					Total Late	Previous Refusals	Previous Non-Refusals	
Size of families								
Total number of families	100,987	4,804	4,392	3,705	229	104	125	---
Percent distribution								
All families	100	100	100	100	100	100	100	100
1 person	30	30	29	29	34	26	40	35
2 persons		33	32	32	34	38	31	47
3 persons	70	17	18	18	16	21	11	10
4 or more persons		20	21	21	16	15	18	8
Age of persons								
Total number of persons	269,277	11,529	10,759	9,196	521	247	274	---
Percent distribution								
All persons	100	100	100	100	100	100	100	100
1-15 years	18	18	19	19	17	15	18	9
15-44 years	38	37	38	39	33	30	36	22
45-64 years	31	31	29	29	32	29	36	48
65 yrs. and older	13	14	14	13	18	26	10	21

1/ Based on data by Census Tracts in United States Censuses of Population and Housing: 1960, Final Report PHC (1) - 104, Part 1.
2/ The results of the second phase of interviewing on a one-third sample of cases not completed in the first phase are weighted by 3. Marginal Distributions = Early Cases + 3 (Late Cases).
3/ Interviewed in November 1960 through April 1961.
4/ Interviewed in May through July 1961.
5/ Based on tabulations by the United States Bureau of the Census of a sample of not-interviewed cases matched with 1960 Census Records.

TABLE 2

Field Status of Dwelling Units, Families and Persons in Estimated Sample; by Size of Family and Age of Persons

	Estimated Sample	Interviewed Early[1]	Interviewed Late[2]			Not Interviewed[2]		
			Total Late	Previous Refusals	Previous Non-Refusals	Total Not Interviewed	Previous Refusals	Previous Non-Refusals
Dwelling units								
Estimated number	4,343	3,329	630	294	336	384	258	126
Percent distribution	100	77	14	6	8	9	6	3
Size of families								
Estimated number of families	4,803	3,705	687	312	375	411	273	138
Percent distributions								
All families	100	77	14	6	8	9	6	3
1 person	100	73	16	6	10	11	6	5
2 persons	100	73	16	7	9	11	7	4
3 persons	100	82	13	8	5	5	3	2
4 or more persons	100	85	12	5	7	3	3	-
Age of persons								
Estimated number of persons	11,527	9,196	1,563	741	822	768	525	243
Percent distributions								
All persons	100	80	13	6	7	7	5	2
Under 15 years	100	85	12	5	7	3	2	1
15-44 years	100	84	12	5	7	4	3	1
45-64 years	100	76	14	6	8	10	7	3
65 yrs. and older	100	73	17	12	5	10	6	4

1/ Interviewed in November 1960 through April 1961.
2/ The results of the second phase of interviewing on a one-third sample of cases not completed in the first phase, weighted by 3, were completed in May through July 1961.

ON THE ACCURACY OF RECORDING OF CONSUMER PANELS

Seymour Sudman, National Opinion Research Center

I. Introduction

There are two major reasons for being interested in the accuracy of recording of consumer panels. Panels have become an increasingly useful source of market research information. Accuracy of response is vital to all surveys. A knowledge of the accuracy of the recording of consumer panels not only makes panel data more useful, but also gives new insights into the general area of survey response accuracy.

This paper is based on research conducted during the past five years with the National Consumer Panel of the Market Research Corporation of America.

Households in the National Consumer Panel report their purchases of a selected group of products on a continuing basis. They use for this purpose self-administered forms called diaries. Diaries for grocery products cover one week's purchase behavior and are mailed in at the end of each week. Diaries for less frequently purchased items, such as clothing and photographic equipment, are kept on a monthly basis.

Households are recruited and trained by personal contact, but additional contacts are normally made by mail. During training, the immediate entry system of recording is taught, i.e., diary keepers are told to record their purchases immediately upon their return from the shopping trip.

The diary forms developed from the accumulated twenty year experience of MRCA in operating panels. As the requirements for information were revised, or as data problems arose, the forms were modified. The research studies described in this paper were not intended to develop a diary de novo, but rather to measure the results obtained with the current forms and to measure the effects of varying these forms.

II. Factors Determining the Accuracy of Recording

First, let us consider the factors determining the accuracy of recording of panel households. Cost considerations have prevented direct measurement of recording accuracy using observational methods.

Indirect measurements using client shipment data are possible, however, after correction for differences in universal definitions.

In observing differences between consumer panel reports and client shipments, it is immediately evident that recording accuracy depends on the characteristics of the product and the techniques used to obtain purchase information. During 1959-1960, seventy-two products were examined for which information was available from National Consumer Panel reports and client shipment data. Many variables were originally considered as being related to accuracy of recording.

The following eight were finally selected for testing:

1. Page in diary
2. Position on page
3. Prominence on page
4. Complexity of entry
5. Frequency of purchase of product class
6. Per cent of product class purchased in chain stores
7. Type of product
8. Convenience of product preparation

Empirical Recording Model. An empirical model was constructed to relate these factors to accuracy of recording. The form of the model was multiplicative. The fitting of the parameters was performed using a simple least squares criterion. Alternative methods of fitting the parameters were considered and tested, but the least squares criterion appeared to be most satisfactory. Computations of the parameters were performed on an IBM 704 Computer. Table 1 shows the computed values of the parameters.

These values agree well with experience in areas other than panel operations. Recording is found to be less accurate for products in the rear of the diary and least accurate for products in the middle, in agreement with the laws of proactive and retroactive inhibition. Accuracy declines as complexity increases. Infrequently purchased products and those not purchased by the housewife are less accurately recorded. The more prominently mentioned products are more accurately recorded as are products at the top of a page.

The values of the parameters have also been tested on fifteen products which were not included in the original model. In no case did the difference between the predicted accuracy of recording and the observed accuracy exceed ten per cent.

III. Factors Influencing Changes in Accuracy of Recording

I shall next describe a series of experiments which measured changes in accuracy of recording resulting from changes in total work load, compensation and training. In each experiment, the Panel was divided into experimental and control groups. The number of diary entries per group was then observed before and during the experiment. The difference due to the experiment was computed as the difference between the changes of the two groups.

Diary Tests. Two major diary tests were conducted. In the first, it was found that recording was more complete in a ledger diary listing the items to be entered than in a journal diary in which purchases were entered in time sequence. Length of diary and position in diary were also found to affect recording, as might have been deduced from the results of the previous section.

A separate monthly diary for miscellaneous products resulted in substantially higher recording for these products. Finally, a weekly check list reminder appeared to have no significant effect on recording. The results of this first test were incorporated into a new diary which was tested for two years and finally adopted for all NCP households.

Table 2 indicates the 2^4 x 3 factorial design with which 432 National Consumer Panel households were tested on 48 different diaries in the period February-April, 1958. Tables 3-6 below show the details for each of the comparisons. Note that in all cases increased recording was considered an improvement. This was confirmed by examination of specific products. Those which were known to be under-recorded, such as non-food and miscellaneous products, had large increases in purchase entries, while those food items which were already recorded completely were unchanged. These conclusions were further verified by a two year experiment with 600 households using a new diary based on the results described above.

Minor Diary Changes. The same methods used for testing major changes in diary format are also used continuously to measure small changes such as moving a product from one page to another, changing its heading, or adding a check box. The following results have generally been observed:

1. Adding or deleting a check box or changing the wording within a check box will not change recording accuracy.
2. Changing a product heading may change the level of recording.
3. Moving a product from one page to another will have the temporary effect of reducing the level of recording.
4. Special reminders result in temporary increases in the level of recording, but not necessarily in long-term changes.

Work Load Changes. Three major changes in work load due to special studies were tested to see whether they influenced the continuing diary keeping. The first was the Menu Study for which households kept records of every meal prepared in the home in a two-week period. The second was a special monthly diary of purchases of gasoline, oil, and auto repairs. The final study was a seven-week study of magazine receivership and television viewing.

None of these three studies had any significant effect on the usual purchase recording behavior. It would appear that, given sufficient compensation, Panel households are willing and able to keep special studies.

Table 7 summarizes the changes in the estimated levels of recording of households in the Menu Study. Table 8 shows the comparison between the Auto Diary households and the control sample by month for the period of January-June, 1961. Table 9 gives the diary entries per household for households keeping media diaries and for all other NCP households for the period January-March, 1959.

Compensation Tests. Households keeping National Consumer Panel diaries received points redeemable in merchandise. Compensation is necessary to keep households cooperating, but the percentage of households who will keep a continuing record of purchases and the accuracy of their recording seems independent of the level of compensation within rather wide limits.

This was seen in a series of four tests, some of which are still in progress. Two of the tests involved new household recruiting, while in the third test, the compensation levels of regular Panel households were changed. In all three of these tests, two compensation levels were used. The lower compensation level was at six-tenths the rate of the higher level.

The fourth compensation test did not involve the regular National Consumer Panel, but was conducted in a midwestern city in which a special consumer panel was recruited. This test market panel kept a short two-page diary and appeared to be sensitive to the level of compensation. The households receiving the higher compensation made twelve per cent more purchase entries per diary than did the households receiving the lower compensation which was at two-thirds the rate of the higher compensation. Table 10 shows the number of purchase entries per diary for low and high compensation households in the test market city.

The National Consumer Panel continuously recruits newly formed households to represent the new household formations among all United States households. These newly formed households are recruited from split-offs of households already in the Panel, i.e., households formed when one or more members of a Panel household move away and establish a new household. Table 11 shows the purchase entries per diary of low and high compensation newly formed households for the period June-September, 1961.

Table 12 shows purchase entries per diary of 150 low and 150 high compensation households during the first thirteen weeks of their Panel membership. Assignments to interviewers were made in groups of ten, of which five households were selected in advance to receive higher compensation and five lower compensation. This design was intended to randomize interviewer effects.

An experiment was begun in January, 1962, which is intended to measure changes in recording behavior when compensation is changed after the household has joined the Panel. The test does not change work load but raises compensation rates by 60 per cent for an experimental group of 100 households and lowers compensation 60 per cent for the second experimental group of 100. Purchase entries per diary of the experimental groups are being compared to purchase entries of control households. Table 13 shows the results through March, 1962.

Special Training. A series of four special contacts with Panel households were tested to determine whether accuracy of recording could be improved by training.

Three of the contacts required the households to return special quizzes related to purchasing. These three quizzes had no important effects on recording. The final contact was a special store shopping record inserted at the front of the regular diary. The objective of this record was to require the household to open the diary after each shopping trip. This shopping record did measurably increase recording, and it was made a part of the new NCP diary.

Table 14 shows the purchase entries per diary of the households receiving the special quizzes for October, 1957-June, 1958 indexed to September,

1957, which was the month prior to the start of the test. The control group for this test was the remainder of the Panel.

Table 15 shows purchase entries per diary of the households who kept the store shopping record compared with the control households. Here, December, 1957, was the base month, since the test ran from January-April, 1958.

TABLE 1

ESTIMATED EFFECTS OF VARIABLES ON ACCURACY OF RECORDING

Diary Page Number	May 1960 All Products
3 - 7	1.00
8 - 26	0.89
27 - 34	0.91
Complexity	
Easy	1.00
Medium	0.91
Complex	0.87
Product Class Purchases per Month in a Panel of 6,000 Households	
Over 2,000	1.00
Under 2,000	0.92
Type of Product	
Food	1.00
Non-Food, Grocery	1.00
Non-Food, Non-Grocery	0.56
Prominence of Entry	
Main	1.00
Subtype	0.86
Position on Diary Page	
Top half of top page	1.00
Bottom half of top page	0.87
Top half of bottom page	0.99
Bottom half of bottom page	0.95
Residual	1.07

TABLE 2

FACTORIAL DESIGN USED IN 1958 DIARY FORMS TEST OF NATIONAL CONSUMER PANEL

		JOURNAL DIARY		LEDGER DIARY	
		Long	Short	Long	Short
With Buying Record Checklist	With Separate Monthly Diary	Sequence a	a	a	a
		b	b	b	b
		c	c	c	c
	Without Separate Monthly Diary	Sequence a	a	a	a
		b	b	b	b
		c	c	c	c
Without Buying Record Checklist	With Separate Monthly Diary	Sequence a	a	a	a
		b	b	b	b
		c	c	c	c
	Without Separate Monthly Diary	Sequence a	a	a	a
		b	b	b	b
		c	c	c	c

TABLE 3

INDICES OF TOTAL PURCHASE ENTRIES BY TYPE OF DIARY, LONG LEDGER (CURRENT DIARY) = 100.0

Type of Diary	Length of Diary		
	Short	Long	Combined
Ledger	110.5	100.0	105.2
Journal	101.6	101.7	101.7
Combined	106.1	100.8	

TABLE 4

INDICES OF PURCHASE ENTRIES OF NCP TEST HOUSEHOLDS FEBRUARY-APRIL, 1958 BY TYPE OF PRODUCT AND POSITION IN DIARY

PRODUCT IN FRONT SECTION=100.0

Type of Product and Diary	Product In 1st Section	Product In 2nd Section	Product In 3rd Section	Product In 2nd or 3rd Sections
Ledger Diary				
Food	100.0	91.6	96.3	93.8
Laundry	100.0	99.6	98.5	99.0
Utility	100.0	84.7	98.0	90.9
Combined	100.0	91.9	96.7	94.3
Journal Diary				
Food	100.0	100.6	100.3	100.4
Laundry	100.0	96.7	89.8	93.1
Utility	100.0	114.9	98.2	105.9
Combined	100.0	101.8	98.8	100.3

TABLE 5

INDICES OF TOTAL PURCHASE ENTRIES OF NCP TEST HOUSEHOLDS FEBRUARY-APRIL, 1958 BY TYPE OF DIARY AND DIARIES WITH AND WITHOUT WEEKLY CHECK LIST REMINDER

DIARIES WITHOUT CHECK LIST REMINDER = 100.0

Type of Diary	With Weekly Check List Reminder	Without Weekly Check List Reminder
Ledger	101.3	100.0
Journal	99.5	100.0
Combined	100.4	100.0

TABLE 6

PURCHASE ENTRIES AND INDICES OF PURCHASE ENTRIES OF MISCELLANEOUS PRODUCTS BY NCP TEST HOUSEHOLDS FEBRUARY-APRIL, 1958 BY WEEKLY VS. MONTHLY DIARIES

WEEKLY DIARY = 100.0

Type of Diary	Purchase entries of Miscellaneous Products	Index
Weekly	350	100.0
Monthly	566	161.7

TABLE 7

CHANGES IN AVERAGE ITEMS RECORDED BY MENU RECEIVERS AND CONTROL HOUSEHOLDS, BY TYPE OF PRODUCT

Items	Net Changes in Items Recorded			Total Items Listed Weekly	Per cent Change
	(1) Menu Receivers	(2) Control	Net Differences (1)-(2)		
A. Total Items					
Pre-Test 1 . . .	-0.28	-0.22	-0.06		0.4%
2 . . .	-0.50	-0.89	+0.39		7.0
3 . . .	+0.84	-0.51	+1.35		8.4
Average	+0.02	-0.54	+0.56	16.1	3.5
S E_d (3 tests combined)					4.0
B. Food Items					
Pre-Test 1 . . .	+0.07	-0.30	+0.37		2.7%
2 . . .	-0.51	-0.83	+0.32		2.4
3 . . .	+0.82	-0.37	+1.19		8.8
Average	+0.13	-0.50	+0.63	13.6	4.6
C. Non-Food Items					
Pre-Test 1 . . .	-0.38	-0.12	-0.26		-10.4%
2 . . .	-0.08	+0.07	-0.15		- 6.0
3 . . .	+0.04	-0.08	+0.12		4.8
Average	-0.14	-0.04	-0.10	2.5	- 4.0

TABLE 8

CODED ENTRIES PER WEEKLY DIARY OF NCP HOUSEHOLDS
KEEPING AUTO DIARY VS. CONTROL GROUP
JANUARY-JUNE, 1961

Month	Coded Entries Per Weekly Diary		
	(1) Auto Diary Households	(2) Control Group	(3) Ratio (1) / (2)
January	8.38	8.32	1.008
February	8.47	8.54	0.991
March	8.27	8.26	1.001
April	8.20	8.08	1.015
May	8.36	8.29	1.010
June	7.90	7.89	1.001
January-June Average	8.20	8.17	1.004

TABLE 9

DIARY ENTRIES/HOUSEHOLD OF MEDIA DIARY RECEIVERS
AND ALL OTHER NCP HOUSEHOLDS
JANUARY-MARCH, 1959

Households	Diary Entries/Household		
	January	February	March
Media diary households	33.7	33.8	32.4
All other NCP households	33.2	33.1	31.6
$S E_d = 0.6$ for monthly comparisons			

TABLE 10

PURCHASE ENTRIES PER DIARY OF LOW AND HIGH COMPENSATION HOUSEHOLDS
CITY:M TEST MARKET - THIRD QUARTER 1961

Period Week Ending	Low Compensation Households Entries per Diary	High Compensation Households Entries per Diary
July 1, 1961	2.30	2.85
July 8	2.33	2.67
July 15	2.47	2.60
July 22	2.40	2.72
July 29	2.07	2.51
August 5	2.34	2.36
August 12	2.14	2.31
August 19	2,30	2.86
August 26	2.37	2.51
September 2	2.22	2.30
September 9	2.35	2.87
September 16	2.40	2.51
September 23	2.42	2.69
Average. . . .	2.32	2.60

SE_d = 0.11 for overall average.

TABLE 11

PURCHASE ENTRIES PER DIARY OF LOW AND HIGH COMPENSATION
NEWLY FORMED NATIONAL CONSUMER PANEL HOUSEHOLDS
JUNE-SEPTEMBER, 1961

Month	Low Compensation Households Entries per Diary	High Compensation Households Entries per Diary
June	9.7	10.2
July	9.1	9.3
August	10.8	9.7
September	9.5	9.6
Average June-September .	9.8	9.7 SE_d=0.9

TABLE 12

PURCHASE ENTRIES PER DIARY OF LOW AND HIGH COMPENSATION HOUSEHOLDS CONTROLLED RECRUITING STUDY, OCTOBER, 1961

FIRST THIRTEEN WEEKS OF RECORDING

Week	Low Compensation Households Entries per Diary	High Compensation Households Entries per Diary
1	15.0	15.3
2	14.6	14.6
3	13.1	13.1
4	13.8	14.6
5	12.7	13.4
6	13.2	12.4
7	13.7	12.9
8	13.6	11.5
9	12.5	14.0
10	15.4	11.6
11	11.4	14.0
12	15.5	12.2
13	14.1	13.3
Average of first 13 weeks	13.8	13.4
SE_d (overall average)	0.6	

TABLE 13

PURCHASE ENTRIES PER DIARY
EXPERIMENTAL VS. CONTROL PANEL HOUSEHOLDS
DECEMBER - MARCH 1962

A. Month	Entries Per Diary: Low Compensation Households Compensation Increased 60%	Low Compensation Households Control	D*
December	13.9	14.1	
January	12.8	13.7	-0.7
February	13.3	14.0	-0.5
March	13.5	14.4	-0.7
Average January-March	13.2	14.0	-0.6

B. Month	Entries Per Diary: High Compensation Households Compensation Reduced 60%	High Compensation Households Control	D*
December	12.0	12.6	
January	12.6	12.0	+1.2
February	12.0	12.6	0.0
March	12.6	12.5	+0.7
Average January-March	12.4	12.4	+0.6
SE_d (overall average)			0.7

$$*D = (X_{Jan.-Mar.} - X_{Dec.}) - (C_{Jan.-Mar.} - C_{Dec.})$$

TABLE 14

INDEX OF PURCHASE ENTRIES PER DIARY
EXPERIMENTAL DIARY REMINDER GROUP
vs.
BALANCE OF NATIONAL CONSUMER PANEL
SEPTEMBER, 1957-JUNE, 1958

September, 1957=100.0

Month	Index of Purchase Entries	
	Diary Reminder Experimental Groups	Balance of NCP
September, 1957	100.0	100.0
October, 1957	105.1	103.1
November, 1957	102.9	104.3
December, 1957 (contacts end)	107.3	103.0
January, 1958	111.3	109.2
February, 1958	108.6	108.8
March, 1958	104.1	107.7
April, 1958	103.0	105.6
May, 1958	103.4	102.9
June, 1958	99.9	99.0

SE_d=2.1

TABLE 15

INDEX OF PURCHASE ENTRIES PER DIARY
OF STORE TRIP RECORD KEEPERS
vs.
CONTROL GROUP
JANUARY-JUNE, 1958

December, 1957=100.0

Month	Index of Purchase Entries	
	Store Trip Record Keepers	Control Groups
December	100.0	100.0
January	108.9	103.1
February	106.7	101.5
March (contacts end)	102.1	97.8
April	104.0	95.4
May	97.2	100.0
June	95.7	92.9

SE_d = 4.2

DISCUSSION

Reuben Cohen, Opinion Research Corporation

When alternative techniques of data collection can be compared on theoretical grounds, agreement on a preferred procedure is not ordinarily a difficult achievement. Thus, probability sampling methods were quickly judged to be superior to sampling methods formerly in use, and there is no occasion for surprise in the fact that each of the sample survey efforts described today is based on a probability sample.

In the broad domain of the methodologist, however, theoretical distinctions are usually more subtle, if they exist at all. Imaginative experimental studies are required to reduce the uncertainties in choice of method. But even the best of experiments, ideally conceived and executed, face the types of questions of interpretation which thread their way through the studies reported today:

(1) Standard of accuracy--against what standard is the outcome of any procedure to be judged?
(2) Level of accuracy--what level is actually required for results to be useful?
(3) Costs--what is the real difference in costs between alternative procedures?

Standard of accuracy

One standard used by Loewenstein and Hochstim is the rate of return or completion rate. Statisticians would agree on 100% completion of interviews or questionnaires as the ideal, and perhaps also on "substantially complete" as a realistic and satisfactory target. In the grey area of substantially complete, however, one would find relatively little consensus, or, at least, different statements of requirements for different purposes.

Hochstim finds ratios of returns for his three strategies ranging from 88% to 93%, and judges each to be satisfactory. The initial implication from Loewenstein is that the 77% rate of completion attained by the first cut-off date was not satisfactory, although the analysis of those interviewed late indicates that adherence to the first cut-off might not have been damaging to the purposes of that study.

But while agreement may be hard to reach on what rate of return is satisfactory, or necessary to a particular purpose, it is at least an objective criterion and we can readily conclude that at given cost, the higher the rate of return the better. Of particular interest, therefore, is the type of procedure designed to improve completion rates (e.g., advanced notification to respondents in the form of the personalized letter described by Hochstim) but which has no substantial effect on costs.

For most methodological experiments there is no standard of accuracy quite so objective as rate of return. Sudman also uses a completeness criterion in that increased recording of purchases is considered an improvement. Apparently satisfactory as a measure of accuracy, at least on an aggregate basis, for the types of product purchases recorded by a consumer panel, it has not proved to be an acceptable standard for other types of inquiries.

Independent sample surveys, or sometimes a complete census, have often served as benchmark measures against which to evaluate specific data collection procedures. Investigators often find themselves in the difficult position of making judgments about the effects of differences in definition of measures used or populations covered in the independent studies, thus, muddling the evaluation of their own data-collection experiments. Nor is it safe to assume that the independent sample survey, or even census, as Bogue points out, is error free.

Hochstim has solved this problem for himself by using identical questionnaires and population definitions for his three different data-collection strategies. While no one technique serves as the standard of accuracy, Hochstim is able, at least, to make judgment about differences among the three, unfettered by the possible effects of definition or other procedural differences.

Level of accuracy

Even more elusive than a definitive standard of accuracy, is an objective statement of requirement for level of accuracy. Properly, the question is not whether the data collection procedure produces the most accurate results possible--but is the level of accuracy suitable to the task at hand? That this is considerably more a matter of opinion than a matter of fact is well illustrated by Bogue's discussion of Census procedures. Most of us design studies with more restricted uses than those to which the Census is put, but we also deal with multi-purpose uses which make level of accuracy requirements difficult to pinpoint.

On occasion our subjective needs are more important than an objective statement of requirements for accuracy. Even if it turns out on complete analysis that the persistance to a higher completion rate, of Loewenstein and her colleagues, failed to change or "improve" the results, the additional effort and costs may be more than justified in other respects. The research team may attribute greater reliability to the data, gain additional confidence in the results, and do a more effective job of communicating conclusions to others.

Costs

In evaluating alternative data collection procedures, some appraisal of costs is inherently necessary. The cost appraisal itself often turns out to be a formidable undertaking, and cost ratios of the type reported by Hochstim are rare and valuable finds.

Frequently we are reduced to speculation about costs, and actions based on tenuous assumptions. Speculations about costs can err, however, and actual cost experiments may turn up surprising results.

VII

STATISTICAL TECHNIQUES FOR UNDERDEVELOPED AREAS

Chairman, Patrick J. Loftus,
Statistical Office of the United Nations

Page

The Post-War Expansion in the Supply of Economic Statistics on the Less-Developed Countries - Abraham Aidenoff, Statistical Office of the United Nations..... 180

Discussion - Stuart A. Rice, Surveys & Research Corporation..... 187

THE POST-WAR EXPANSION IN THE SUPPLY OF ECONOMIC STATISTICS ON THE LESS-DEVELOPED COUNTRIES

Abraham Aidenoff, Statistical Office of the United Nations

Though the supply of economic statistics relating to the less developed countries of the world is still far from adequate, during the post war period - especially the last 10 years or so - the governments of these countries have markedly increased the extent of these series of data. This expansion in the supply of economic statistics resulted primarily from new requirements for data. Prior to World War II, economic statistics were wanted by the governments of the less developed countries for such purposes as measuring sources of revenue and administering the day-to-day affairs of various departments. Since the early 1950s the efforts to develope economically have furnished the major source of requirements for economic statistics. The consequences of this shift in the character of the demands for data have been not only substantial expansion in the supply of economic statistics but also marked changes in the organizational and other arrangements for gathering and compiling these data.

In this paper we will deal with the subject of economic statistics relating to the less developed countries of the world from the points of vantage that were just mentioned. The countries of Latin America, Asia and Africa will be treated as less developed. We will discuss first the changes, post-war, in the supply and character of selected economic statistics in the light of the requirements for data for purposes of furthering and assessing economic development. We will then go on to examining the changes in the arrangements for gathering and compiling economic statistics that accompanied the extension in these data.

The needs for data on the structure and interrelationships of an economy in devising plans and programmes for its development have furnished a good part of the motivation behind the economic censuses and similar inventories taken during the post-war period. The same type of requirements for data have stimulated basic inquiries into the financing of certain aspects of production, especially capital investment, into the sources of demand for goods and sources, and into the interrelationships between production, incomes, expenditures and savings.

Even broad and sketchy efforts to further development require data on the character of the economy involved. Information is needed, for example, on the contribution to production, incomes, exports, on the employment of labour and capital and on the demand for key materials and other intermediate goods of sectors of the economy such as agriculture and the various mining and manufacturing industries. Similar but, perhaps, less extensive, data are wanted on construction and transportation and similar public utilities. Such data are essential to delineating the parts of the economy on which the efforts to develope should focus, to defining the policies and instruments which might further these efforts, and to maintaining the necessary balance between the various sectors of economic activity. For these purposes, information on the supply of key goods and services from domestic production needs to be supplemented with data on the details of imports since in the case of the less developed countries, in particular, imports are a key source of new capital equipment and substitutions of domestic production for imports furnish possible avenues for economic expansion. On the side of the demand for goods and services, data are required concerning dimensions of personal consumption, government expenditures, fixed capital formation and exports. The relative magnitudes of personal expenditures, personal and other incomes and domestic savings, on the one hand, and domestic savings, sources of external finance and capital formation, on the other, also provide needed intelligence. Information on the sources of demand for goods and sources is essential to charting realistic expansion in domestic production, which will avoid the pitfalls of deficiencies or surpluses in the supply of goods and services and the consequent inflation or deflation. Data on the interrelationships between incomes, current expenditures, capital formation and sources of capital finance are also needed for these purposes.

It is no wonder, therefore that, as part of their efforts to develope economically, the governments of the Southern Hemisphere have become forcefully aware of the importance of data on the basic characteristics of their economies. Even where they have not engaged in systematic economic development, turning to the World Bank, other international agencies, or highly industrialized countries for financial and other assistance has usually resulted in requests for such data as part of the application and justification for the aid. The works of various international organizations and bodies on the problems and means of economic development, as well as on statistics as such, has also brought home to these governments the requirements for basic economic data. It should be emphasized, however, that in a significant number of countries, awareness of the need for such statistics has not yet resulted in the taking of the pertinent statistical inquiries. This has been due, among other reasons, to the limited funds and resources for statistical work, the lack of experience and the difficulties involved in taking these inquiries. Nevertheless, concern with economic development has furnished a drive, which was missing before World War II, for engaging in the collection and compilation of economic statistics. Governments could hardly be expected to spend much effort in gathering data which were, to a considerable extent, of academic interest before.

The first six tables set out in the annex to this paper illustrate, in the case of selected countries in Latin America, Asia and Africa, the increasing prevalence during the post-war period, of certain inquiries, statistical series or sets of accounts yielding basic data on the character and structure of the economy.

Table 1 is devoted to comparing the number of countries which gathered data, in the 1940 and 1950 censuses of population, on the industrial

distribution of the economically active populati tion. This kind of data furnishes a useful measure of the relative significance of the various kinds of economic activity which, in the case of the less developed countries, is not beset with as serious problems of concept or collection as some other broad measures of the structure of production. This is the case because, in the less developed countries, the household is the site of much production, both commercial and subsistence, and productive activity is frequently not clearly segregated from other human activities. And, censuses of population of course require enumeration of the characteristics of each individual, whether or not he is working on a farm, in a mine, factory, shop, store or some other institutionalized productive unit. The industrial distribution of the economically active population, taken in conjunction with its occupational distribution, also furnishes a useful inventory of the kinds of labour available.

It is evident from Table 1 that a marked increase took place from the 1940 round of censuses of population to the 1950 round in the number of less developed countries gathering data on the industrial distribution of the economically active. This increase was largely due to the much higher proportion of the countries taking population censuses which gathered these data. For example, among the Latin American countries analyzed, only 9 of the 16 countries participating in the 1940 round of censuses gathered data on the industrial distribution of the economically active, whereas this was the case for all of the countries taking part in the 1950 round. In the case of the Asian countries analyzed, the comparable figures were 1 of 10 in 1940 round and 10 of 10 in the 1950 round. Preliminary study also shows that all of the Latin American and Asian countries participating in the 1960 round of censuses of population have gathered, or propose to gather, data on the industrial distribution of the economically active.

Censuses of agriculture furnish inventories of the predominant economic activity in the less developed countries. Nevertheless, only 7 of 20 Latin American countries and 2 of the 18 Asian countries took censuses of agriculture during the 1930 round of these inquiries. These small numbers are especially surprising in view of the significance, in a number of these countries, of taxes on agricultural land or output as a source of governmental revenue. Under the impetus of the data required for efforts to develope, the comparable figures for the 1950 round of censuses were 19 and 7, respectively. Moreover, many more Asian and African countries are participating in the 1960 round of censuses of agriculture than in the 1950 round.

Mining, manufacturing, construction and the production of electricity and gas are a major focus of attention in programmes for economic development. Economic development often involves markedly expanding these activities shifting underemployed agricultural labour to the industrial sector, diversifying and extending the scope of manufacturing, and substantial fixed capital formation in these industries. It is not too surprising, therefore, that all of the Latin American countries analyzed and 17 of the 18 Asian countries are participating in the 1963 World Programme of industrial censuses and similar basic inventories. Before 1948, only 6 of the Latin American countries and 2 of the Asian countries had taken such inquiries. Even half of the African countries studied are taking part in the 1963 Programme. The comparable figures for the 1953 round of these censuses were 11, 6 and 8.

Table 4, which deals with statistics of external trade, illustrates that compilation was as common in the pre-war as the post-war period in the case of data on transactions which have always been a major source of government revenue and the subject of detailed government control. Information on imports and exports were, and still are, gathered, of course, as part of the administration of the government customs. Though the raw data were as available pre-war as post war, because of post-war requirements for these data, marked changes have taken place in the way in which these figures are compiled. Almost all of the countries analyzed have added, detailed and summary tabulations of imports and exports according to classifications which are economically meaningful, usually modeled on the Standard International Trade Classification, to the detailed tabulation according to categories of varying custom duties. The demands for these data arising out of economic analysis and policy making have also resulted in much more emphasis, post-war than pre-war, on the compilation of figures of quantities, as well as values, for exports and imports.

The substantial differences, post-war and pre-war, in the availability of statements of the balance of payments, in contrast to the situation in the case of external trade, point to the stimulants to gathering economic statistics furnished by the efforts of the countries studied to develop economically. As compared to the 20 Latin American, 14 Asian and 11 African countries which had statements of this balance of payments in 1960, only 7, 4 and 2, respectively, of these countries had such statements in 1938. As we all know, the composition of the balance of payments is of fundamental importance to the less developed countries, all faced with finding the foreign exchange to finance imports of much of the capital goods strategic for economic development.

A set of national accounts on production expenditure and income furnishes an over-all view of the character and structural interrelationships of an economy. Such accounts have therefore come to be considered essential tools in plans and programmes for furthering economic growth. As a consequence, by 1960, all of the Latin American countries studied, 14 of those in Asia and 11 in Africa had devised sets of national accounts. It must be emphasized that the reliability of the figures presented in a number of the accounts, as well as the degree of articulation of these accounts, left much to be desired, since the basis for a set of adequate national accounts is

a highly developed body of economic statistics. Nevertheless, the situation, in 1960, represented a marked improvement over that existing during the early part of the post-war period. By 1948, only 15, 6 and 4, respectively, of these countries had tackled the work of compiling such accounts.

In addition to data on the character and structure of their economy, countries, in dealing with economic development, require data on the trends, in terms of quantities and prices, for such aspects of their economy as the various kinds of production, the sources of expenditures on goods and services, and external trade. An obvious part of efforts to raise the level of economic development is ascertaining the extent to which goals have been reached and detecting current disbalances and other problems requiring attention.

Some of the economic statistics discussed before, such as sets of national accounts, statements of the balance of payments and data on external trade, are as essential in measuring current developments as in delineating basic features of an economy. The last six tables of the annex to this paper are devoted to comparing the availability, post-war and pre-war, of some of the additional series of statistics which are required for assessing current developments. The figures presented in these tables suggest that in the case of annual or more frequent economic statistics, the less developed countries have not yet attained the stage that they have reached in the case of benchmark data. This seems to be due, at least in part, to the less rapid post-war expansion in the more frequent statistics.

Set out in Table 7 is a comparison, pre-war and post-war, of the number of the countries which gathered sufficient annual data on the output of important crops to construct index numbers of food production. During the post-war period utilized, twelve of the Latin American countries met this criterion; an addition of 4 over the pre-war period. The figures for Asian and African countries are 12 and 10 and 6 and 5, respectively.

The increase during the post-war period in the number of less developed countries taking annual or similar industrial inquiries is greater than that for countries gathering data on the output of crops. For example, 16 Latin American countries took such surveys during the years, 1956-1960, as compared to 2 before 1948. The contrast between the post-war and pre-war periods is about the same in the case of the Asian and African countries. Nevertheless, rather fewer countries took annual surveys of the industrial sector than censuses or similar benchmark inquiries. For example, 4 Latin American countries took the former type of surveys during the period, 1950-1955, but 11 of these countries conducted industrial censuses around 1953..

The less developed countries made even less progress, post-war, in compiling annual or more frequent index numbers of manufacturing production and employment than in taking annual industrial surveys. Table 9 shows, for example, that 10 Latin American countries compiled indexes of production during the period, 1954-1958. Three of these countries produced the series before World War II. Comparable figures for Asia and Africa are 7 and 2, each. The dimensions of the increase, between the pre-war and post-war periods, in the case of index numbers of manufacturing employment was similar to that for index numbers of manufacturing production.

Somewhat more less developed countries are compiling series of index numbers on wholesale prices than on manufacturing production or employment. However, these countries have made little, if any progress, post-war, in this phase of their statistical work. This situation may be indicative of the much lesser attention devoted in work on economic development to dealing with price aspects than with the "real" aspects.

Contrary to the situation in the case of index numbers of wholesale prices, it is evident from Table 12 that index numbers of cost of living have been the subject of considerable attention on the part of the less developed countries. Perhaps this reflects the need for indexes of cost of living to ascertain the impact of development efforts on "real" levels of living.

I hope that the preceding review of the post-war expansion in the availability of selected economic statistics, has not left the impression that the less developed countries now have adequate data on their economies. As any of us who have worked on, or utilized, these data know, this is far from the case. The lack of some of the required economic statistics of a general type-annual or more frequent and benchmark has been indicated previously. The discussion has not dealt, however, with the availability of other general economic data - for example, on the distributive trade, transportation and other services - much of which is less plentiful than the series that have been discussed. Further, attention has not been devoted to the gaps of coverage and the deficiencies of detail, definition, unreliability and lateness in the data that are gathered and compiled. However, it is clear from our brief review that during the last 10 years or so, the governments of less developed countries have, on the whole, made considerable progress in widening the collection and compilation of the data required on their economies.

In many less developed countries, the introduction of the new series of data on the economy has gone, hand-in-hand, with the evolution of new arrangements for gathering and compiling economic statistics. Unlike the bulk of the economic data gathered pre-war, most of the new series could not be obtained as a by-product of governmental administrative activities. Instead, the data had to be gathered directly from businesses or households, through inquiries designed especially for this purpose. Further, in order to plan and take these inquiries effectively, it was essential to make use of the concepts, methods and techniques of statistics rather than those of governmental supervision, control or accounting.

This involved training and recruiting a corps of statistical personnel and accumulating and effectively retaining experience in such work. The organizational bases for tabulating the results of these inquiries, as well as other economic data, in ways which were most useful for statistical and economic analysis involved the same pre-requisites.

These circumstances, as well as some considerations have led the governments of many less developed countries to establish and foster special organizations and corps of personnel for statistical work. In a number of Asian countries, for example, new departments and offices have been founded which were devoted to the collection and compilation of economic and other types of statistics. And, increasingly, statisticians and other statistical personnel have been recognized as distinct corps of civil servants. The same events have already taken place in some African countries. In other countries of Africa, the setting up of statistical organizations and the development of statistical personnel has just begun. In Latin American countries where continuing statistical organizations - essentially for census taking - did not exist before, permanent agencies for statistical work have also been established. Where census offices already existed, the functions of these agencies have been extended to cover the taking of many other kinds of statistical inquiries and the compilation of other types of statistical series, including data derived from governmental administrative records. In the countries of all three regions, government units, agencies and departments which are not specialized in statistical work now play a much less important role in collecting and compiling economic statistics than they did pre-war, or during the early part of the post-war period.

The post-war emphasis on the use of economic statistics in planning and assessing economic development led to establishing special statistical organs - in particular, central statistical offices - because of some other considerations in addition to those mentioned before. As we noted earlier, arraying the various kinds of economic statistics into integrated national accounts materially facilitates the use of the data in work on the problems of economic development. Even where formal integration does not take place, it is essential to have data on each of the important aspects of the economy and to examine and analyze the various economic statistics in relation to one another. In order to meet these requirements, the statistical apparatus, activities and plans of a country have to be viewed as a whole and the concepts, classifications, definitions and techniques utilized in gathering and compiling the various economic statistics need to be co-ordinated with one another. Such an approach to the statistical activities of a government has generally involved setting up an apparatus for planning and co-ordinating these activities in the light of both the requirements for data and the circumstances under which the data have to be gathered and compiled.

These considerations, in addition to the prerequisites mentioned earlier, have led a number of Latin American, Asian and African countries to concentrate almost all their statistical work in one central statistical organization. In the case of other countries in these regions, where statistical work is still dispersed among a number of specialized units, co-ordinating statistical units have been established, which also often compile the national accounts for the country and other over-all economic indicators. In a number of instances, the central statistical or co-ordinating agency is part of either the economic development organization or a ministry which is also responsible for economic planning and development. In a few countries, programmes for statistics have been made part of articulated plans for economic development. In any case, in most of the countries, clearly defined means and apparatus have been, or are being, established for planning and carrying out of the statistical activities of the government so that the requirements for data for purposes of work on economic development will be satisfied.

The organizational arrangements for statistical work that are evolving in the less developed countries are promising for more rapid progress in extending and improving the available economic statistics in the near future than in the past. These arrangements are conducive to the establishment of a growing corp of trained and experienced statistical personnel. Explicit and systematic study of the gaps and inconsistencies in the supply of available economic statistics in the light of the pressing requirements for these data should stimulate the rectification of these deficiencies. The increasing concerted use of the available figures in the various tasks of planning, furthering and assessing economic development should furnish real incentives for improving the supply co-ordination and quality of the economic statistics gathered and compiled and for providing the funds and resources for these purposes.

ANNEX

Table 1. Data on the Industrial Classification of the Economically Active Population

Region	Number of Countries Analyzed	Number of Countries Compiling Data	
		During the 1940 Censuses of Population	During the 1950 Censuses of Population
Latin America	20	9	19
Asia, East and Southeast	18	1	10
Africa	36	2	11

Table 2. Censuses of Agriculture

Region	Number of Countries Analyzed	Number of Countries Taking the Census	
		During 1930 Round	During 1950 Round
Latin America	20	7	19
Asia, East and Southeast	18	2	7
Africa	36	...	17

Table 3. Industrial Censuses or Similar Inquiries

Region	Number of Countries Analyzed	Number of Countries Taking the Inquiry		
		Before 1948	During 1953 Programme	During 1963 Programme
Latin America	20	6	11	20
Asia, East and Southeast	18	2	6	17
Africa	36	2	8	18

Table 4. Data on External Trade

Region	Number of Areas Analyzed	Number of Areas Compiling Data	
		For 1938	For 1948 and thereafter
Latin America	20	20	20
Asia, East and Southeast	18	17	18
Africa	23 1/	22	23

1/ British East Africa - 4 countries, former French Equitorial Africa - 3 countries, former French East Africa - 8 countries; are each counted as one.

Table 5. Statements of the Balance of Payments

Region	Number of Areas Analyzed	Number of Countries Having Statements		
		In 1938	In 1948	In 1960
Latin America	20	7	15	20
Asia, East and Southeast	18	4	6	14
Africa	36	2	4	11

Table 6. National Accounts on Product, Expenditure and Income

Region	Number of Countries Analyzed	Number of Countries Compiling the Accounts	
		In 1948 or earlier	In 1959-1960
Latin America	20	3	12
Asia, East and Southeast	18	3	13
Africa	36	2	14

Table 7. Data on Output of Agricultural Commodities for Index Numbers of Food Production

Region	Number of Countries Analyzed	Number of Countries Furnishing sufficient Data to Compile the Index Numbers on:	
		Pre-war Base (1934-1938=100)	Post-war Base (1952/1953-1956/1957=100)
Latin America	20	8	12
Asia, East and Southeast	18	10	12
Africa	36	5	6

Table 8. Annual or Somewhat Less Frequent Industrial Inquiries

Region	Number of Countries Analyzed	Number of Countries taking Inquiries		
		Before 1948	1950-1955	1956-1960
Latin America	20	2	4	16
Asia, East and Southeast	18	0	2	9
Africa	36	2	3	15

Table 9. Index Numbers of Manufacturing Production

Region	Number of Countries Analyzed	Number of Countries Compiling the Index Numbers		
		Pre-war	During 1948-1953	During 1954-1958
Latin America	20	3	8	10
Asia, East and Southeast	18	2	6	7
Africa	36	2	4	7

Table 10. Index Numbers of Manufacturing Employment

Region	Number of Countries Analyzed	Number of Countries Compiling the Index Numbers	
		Around 1948	1955-1960
Latin America	20	6	10
Asia, East and Southeast	18	2	5
Africa	36	1	7

Table 11. Index Numbers of Wholesale (Producer) Prices

Region	Number of Countries Analyzed	Number of Countries Compiling the Index Numbers	
		Around 1948	Around 1960
Latin America	20	13	13
Asia, East and Southeast	18	10	8
Africa	36	5	9

Table 12. Index Numbers of Cost of Living

Region	Number of Countries Analyzed	Number of Countries Compiling the Index Numbers	
		Around 1948	Around 1960
Latin America	20	17	19
Asia, East and Southeast	18	14	14
Africa	36	14	24

DISCUSSION

Stuart A. Rice, Surveys & Research Corporation

My comments refer to the difficulties and problems of statistical activities in underdeveloped areas, some of which have been mentioned in the papers just presented, especially by Dedrick; rather than to their results as summarized by Mauldin and Aidenoff. The results are often impaired by difficulties of which statisticians from other areas are unaware. Census taking offers many illustrations. The net underenumeration of the population of the United States in the 1960 census was estimated at a meeting yesterday to be between 3 and 3½ million. It is evident that difficulties in obtaining an accurate count still exist; but a larger number have undoubtedly been overcome by experience and research during the nine decades since the historic report by General Francis A. Walker on the problems and deficiencies of the census of 1870. We cannot be too complaining about inaccuracies in census figures in areas lacking the long experience and public habituation to census inquiries of the United States; but we should not be unaware of their existence and the reasons for them.

When statisticians from advanced nations are asked to advise in areas of lesser development, frustration may result from barriers of communication. These are often augmented by differences in social institutions and attitudes. Perhaps we try to inculcate in the indigenous statistician the idea--at variance with his training, which has been one of learning by rote--that he should think for himself. He is inspired by this new idea, only to be criticized for seeking when in the field to improve on the exact specifications of a carefully designed sample survey.

The barriers to be overcome in extending the use of statistical techniques to underdeveloped areas are thus not merely linguistic but broadly sociological and cultural. Family organization, relationships between employers and employees, even such seemingly definite conceptions as age, may be viewed and recorded in statistical surveys from standpoints very different from our own. Age in Korea traditionally refers to the number of separate calendar years within which one has lived (complicated further by a lunar calendar) and not, as among ourselves, to the number of years of 12 months which have elapsed since birth.

Again, what meanings can be attached to "labor force" and "national income" when a national economy is based upon small agricultural holdings, consumption in the household of what is produced, and a considerable amount of barter? Figures on these subjects derived by the methods of imputation used in more advanced economies are hazardous for international comparisons. Such problems are not uniquely those of underdeveloped countries. They have earlier been encountered in the more advanced nations which have found solutions consistent with their own economic order and institutional patterns.

There has been a recent shift of attention among international organizations from the promotion of statistical comparability between countries to efforts directed at indigenous statistical development. The growth of indigenous statistical systems, adapted in each case to the special needs of the countries concerned, seems in need of greater attention.

VIII

ANALYTIC STATISTICS IN SURVEYS

Chairman, Eli S. Marks, Case Institute of Technology

Page

Variances for Indexes from Complex Samples - Leslie Kish, University of Michigan..... 190

A Test of Homogeneity for a Stratified Sample - Tomas Garza-Hernandez, C-E-I-R de Mexico and Philip J. McCarthy, Cornell University..... 200

Multiple Frame Surveys - H. O. Hartley, Iowa State University.... 203

Sampling for Weekly Time Series - Max A. Bershad and Harold Nisselson, U. S. Bureau of the Census..... 207

VARIANCES FOR INDEXES FROM COMPLEX SAMPLES *

Leslie Kish, Survey Research Center, University of Michigan

Summary

Methods of computing variances were found for several statistics of importance in economic and social surveys. These statistics are based on the ratio mean (score) $r = y/x$, where x is commonly the variable sample size of complex (stratified clustered) samples. One set of these statistics involves the weighted sum $\sum^{J} w_j r_j$ for any real w_j. Another set is based on the relatives $R_1 = r_1/r_o = (y_1/y_1)/(y_o/x_o)$, the ratio of current (1) score to the base (o) score. From these relatives the indexes $I_1 = \sum^{J} R_{1j}$ are constructed and then changes in the index $I_2 - I_1$. We develop computing formulas for the variances of these statistics, including some good simple approximations. These methods are applied to a series of data from the Surveys of Consumer Expectations. A large group of empirical results give interesting and useful evidence on the behavior of those statistics, with wide implications for the design of surveys to measure economic and social indicators.

(1) Variances for Linear Combinations of Ratio Means

The use of sample surveys for collecting social and economic data is increasing in scope and some important indexes are now--and more will be in time--based on sample surveys. Many of these resemble in the essentials of sample design the Surveys of Consumer Expectations [2] that provided both the stimulation and the data for our empirical investigations. Our principal aims were empirical: to learn something about the magnitudes and sources of sampling variations affecting the principal economic indicators. The surveys are based on a complex multi-stage sample, from which the scores r are computed as the ratios of two random variables y and x:

$$r = y/x = \sum_{h}^{H} y_h \Big/ \sum_{h}^{H} x_h = \sum_{h}^{H}\sum_{\alpha}^{a_h} y_{h\alpha} \Big/ \sum_{h}^{H}\sum_{\alpha}^{a_h} x_{h\alpha} \qquad (1)$$

Typically y is the sum of some characteristic in the sample and an estimate (usually unbiased) of the population value Y, except for a constant factor--such as f, the sampling fraction. The random variable x in the denominator is here (and usually) the sample size; in general it estimates the population value X in a form similar to the numerator. The sample is selected in H primary strata and a_h denotes number of primary selections from the h-th stratum. Our data come from 45 strata representing the U.S.A. and $a_h = 2$ primary selections (mostly counties) from each stratum. The quantities $y_{h\alpha}$ and $x_{h\alpha}$ represent sample totals for the α-th primary selection in the h-th stratum; $y_{h\alpha}$ is the total score for an attitude and $x_{h\alpha}$ the sample size; the quantities y_h and x_h are stratum totals [4].

The variance of the sample ratio r can be estimated [1, 4, 8] usefully in large samples by

$$\mathrm{var}(r) \doteq x^{-2}[\mathrm{var}(y) + r^2\,\mathrm{var}(x) - 2r\,\mathrm{cov}(y,x)] \qquad (2)$$

We shall use the convenient computing forms [4] for the variance of a score r

$$\mathrm{var}(r) \doteq x^{-2} \sum^{H} dz_h^2 = \sum^{H} [dz_h/x]^2 \qquad (3)$$

where

$$dz_h^2 = \frac{a_h}{a_h - 1} \sum^{a_h} [Dy_h - rDy_h]^2$$

$$= \frac{a_h}{a_h - 1} \sum_{\alpha}^{a_h} [(y_{h\alpha} - y_h/a_h) - r(x_{h\alpha} - x_h/a_h)]^2 \qquad (4)$$

when abbreviated to the basic units of deviation. Often, as in our sample, two primary selections (a and b) are drawn from each stratum; and with $a_h = 2$ the basic computing unit takes the form:

$$dz_h = dy - rdx = [y_{ha} - y_{hb}) - r(x_{ha} - x_{hb})] \qquad (4')$$

This is quite common in survey practice and Keyfitz [3] called attention to the wide usefulness and simplicity of that design. But I believe the essential simplicity and flexibility is in the computing forms (3) and (4), which can also be used for more than two selections with either electronic or desk computers. Similar to the variance, the covariance of two scores r_j and r_k, based on samples from the same primary selections, is:

$$\mathrm{cov}(r_j, r_k) = \sum_{h}^{H} (dz_{jh}/x_j)(dz_{jh}/x_k)$$

$$= (x_j x_k)^{-1} \sum_{h}^{H} dz_{jh} dz_{kh} \qquad (3')$$

These forms permit readily the computation of variances for the linear combination $\sum^{J} w_j r_j$ of several (J) scores r_j, where the w_j are constants as needed: (5)

$$\mathrm{var}\Big(\sum_{j}^{J} w_j r_j\Big) = \sum_{j}^{J} w_j^2\,\mathrm{var}(r_j) + 2\sum_{j<k}^{J} w_j w_k\,\mathrm{cov}(r_j, r_k)$$

$$\doteq \sum_{j}^{J}\sum_{h}^{H} \Big(w_j \frac{dz_{jh}}{x_j}\Big)^2 + 2\sum_{j<k}^{J}\sum_{h}^{H} \Big(w_j \frac{dz_{jh}}{x_j}\Big)\Big(w_k \frac{dz_{kh}}{x_k}\Big)$$

$$\doteq \sum_{h}^{H} \Big[\sum_{j}^{J} w_j \frac{dz_{jh}}{x_j}\Big]^2 = \sum_{h}^{H} dv_h^2 \qquad (5')$$

* Research supported by grant G-7571 for "Analytical Statistics for Complex Samples" from the National Science Foundation. I am especially grateful for the collaboration of Mr. R.K. Pillai who took complete charge of the mass of complex computations.

The first line follows from basic theorems about linear combinations of random variables. In the second line we reverse the order of summation to obtain the third line readily, because of the simple additivity of the within stratum computing forms.

An example of (5) is the change of score $(r_2 - r_1)$ from period 1 to period 2; here J = 2 with $w_2 = 1$ and $w_1 = -1$:

$$\text{var}(r_2 - r_1) \doteq \sum^{H}(\frac{dz_{2h}}{x_2} - \frac{dz_{1h}}{x_1})^2 = x_2^{-2}\sum^{H} dz_{2h}^2 + x_1^{-2}\sum^{H} dz_{1h} - 2(x_2x_1)^{-1}\sum^{H} dz_{2h}dz_{1h} \quad (6)$$

Another example is the simple summed score, $\sum^{J} r_j = \sum^{J} y_j/x$, when all $w_j = 1$, all $x_j = x$ and $y^*_{h\alpha} = \sum_{j}^{J} y_{h\alpha j}$

$$\text{var}(\sum^{J} y_j/x) \doteq x^{-2}\sum^{H}(\sum^{J} dz_{jh})^2 = x^{-2}\sum^{H}(\sum^{J} dy_{jh} - Jrdx_h)^2 = x^{-2}\sum^{H}(dy_h - Jrdx_h)^2 \quad (7)$$

2. The Variances of Relatives and Indexes

The relative for any current year is the ratio of the score for that year to the score for the base year $R_1 = r_1/r_o$.

$$R_1 = r_1/r_o = (y_1/x_1)/(y_o/x_o) = (y_1x_o)/(y_ox_1) \quad (8)$$

The requirement that r_o, hence y_o, be safely far from zero can usually be met and the base period score is often set arbitrarily at 100 (or 1). The values of x in the denominators of all scores must be such that compared to sampling variability we can disregard the possibilities of values near zero. This requirement is met readily when, as in our sample the x values denote sample sizes under good control [5]. We also use the difference of two relatives $(R_2 - R_1)$.

Finally, we also investigate indexes of the type

$$I_1 = \sum^{J} R_{1j} = \sum^{J} r_{1j}/r_{oj} \text{ and } I_2 - I_1 = \sum^{J} R_{2j} - \sum^{J} R_{1j} \quad (9)$$

the difference of two indexes. The difference typically indicates a change between two periods (2 and 1) in the indexes I_2 and I_1; this can refer either to the entire sample (and population) or to a subclass. But the same treatment can also accommodate other kinds of differences, such as the comparison of indexes for two subclasses in the same period.

The index is the sum of J relatives, R_{1j}, and combining different relatives results in different indexes. Each relative, R_{1j}, consists of the current score r_{1j} standardized by taking it as the ratio to the score, r_{oj}, for the base period "0". If before summation the relatives are multiplied by weights, these can be conveniently assumed incorporated either in the base score r_{oj} or (if the weights are subject to periodic changes) in the current score r_{1j}.

These standard economic measures are well known and frequently used. But they need to be accommodated to a sampling approach when the data come from sample survey data, particularly from a complex sample design of the kind commonly used for large-scale socio-economic surveys. We provide methods for computing variances for these indicators and then use them for the empirical investigation of a large and important source of data. Further problem about optimum weights and methods of combining the indicators must be left to other investigations.

If we treat $R_1 = r_1/r_o$ simply as the ratio of two random variables, the variance of which may be estimated with the same formula (2) that we use for $r = y/x$, then we have

$$\text{var}(R_1) \doteq \frac{1}{r_o^2}[\text{var}(r_1) + R_1^2\,\text{var}(r_o) - 2R_1\,\text{cov}(r_1, r_o)] \quad (10)$$

This brings our only serious new theoretical problems. We can and must be satisfied with a large sample approximation, because most surveys are based on large enough samples and this permits the use of two approaches. We note that the derivations [1, 4] used for obtaining (10) need few restrictions. Especially the variances of r_1 and r_o can be mean-square errors and small biases will not destroy these relationships; we found these biases of r_1 and r_o to be trivial or absent. We also investigated the bias in the ratio $R_1 = r_1/r_o$, and found it to be real, but negligibly small for our purposes.

Another approach to this problem is through methods sometimes called the "propagation of errors"; see for example Rao [7]. We must depart from his classical presentation which uses terms $\sigma_{ij}/\sqrt{n}$, assuming simple random sampling. This was generalized to the covariance terms of a large sample of complex design by Mr.'s R.K. Pillai and N.K. Namboodiri--as yet in rough draft form. The approximation results in a relatively simple form:

$$\text{var}(R_1) \doteq R_1^2\sum_{h}^{H}[(\frac{dy_1}{y_1} - \frac{dx_1}{x_1}) - (\frac{dy_o}{y_o} - \frac{dx_o}{x_o})]_h^2 \quad (10')$$

A form of this, a "double-ratio sample" for a random sample is attributed by Yates [8, p. 343]

to Keyfitz, and mentioned briefly by Hansen [1, Vol. I, p. 514]. This can be generalized for $U/V = \prod_j u_j / \prod_j v_j$ as:

$$\mathrm{var}(U/V) \doteq (U/V)^2 \sum_h^H [\sum_j^J \frac{du_{hj}}{u_j} - \sum_j^J \frac{dv_{hj}}{v_j}]^2 \qquad (11)$$

but the validity of the approximation must be checked empirically in each case.

The computation of indexes can be facilitated with a different computing unit; which bring the R_1 inside the brackets of (10'). In addition to some computational conveniences this also avoids division by y_1, hence the unnecessary assumption that $y_1 \neq 0$. Thus:

$$e_{1h} = \frac{dz_{1h}}{r_o x_1} = \left(\frac{dy_{1h} - r_1 dx_{1h}}{r_o x_1}\right) \text{ and } e_{oh} = \frac{dz_{oh}}{y_o} = \frac{dy_{oh} - r_o dx_{oh}}{y_o} \qquad (12)$$

because $r_o x_o = y_o$. With these units we get (10') into the brief form

$$\mathrm{var}(R_1) \doteq \sum_h^H (e_{1h} - R_1 e_{oh})^2 \qquad (10'')$$

We may also express in these e_h units the variances of scores

$$\mathrm{var}(r_1) \doteq r_o^2 \sum_h^H e_h^2 \qquad (3')$$

and the difference of two scores to the same base r_o

$$\mathrm{var}(r_2 - r_1) \doteq r_o^2 \sum_h^H (e_{2h} - e_{1h})^2 \qquad (6')$$

From simple considerations of symmetry we also see that the variance of the product $(r_1 r_2)$, which has a positive covariance term, can be computed as

$$\mathrm{var}(r_1 r_2) \doteq \sum_h^H [\frac{1}{x_1 x_2}(y_2 dz_{1h} + y_1 dz_{2h})]^2 \qquad (10''')$$

This can be generalized to the product of more than two ratios insofar as this would be justified by its distribution. Furthermore, it could also be summed in the manner of formulas (14) and (15) below.

The useful simplicity of the e_h units appears soon when we need variances of linear combinations of of the relatives R_j. The difference of two relatives measures its change between two periods and its variance appears term by term as: (13)

$$\mathrm{var}(R_2 - R_1) = \mathrm{var}(R_2) + \mathrm{var}(R_1) - 2\mathrm{cov}(R_2, R_1)$$

$$\doteq \sum_h^H [(e_{2h} - e_{1h}) - (R_2 - R_1)e_{oh}]^2 \qquad (13')$$

$$= \sum_h^H (e_{2h} - e_{1h})^2 + (R_2 - R_1)^2 - 2(R_2 - R_1)\sum_h^H e_{oh}(e_{2h} - e_{1h}) \qquad (13'')$$

$$= r_o^{-2}\mathrm{var}(r_2 - r_1) + (R_2 - R_1)^2 r_o^{-2}\mathrm{var}(r_o) - 2(R_2 - R_1)r_o^{-2}[\mathrm{cov}(r_o, r_2) - \mathrm{cov}(r_o, r_1)] \qquad (13''')$$

The last expression can also be obtained immediately by considering $(R_2 - R_1) = (r_2 - r_1)/r_o$ as the ratio of two random variables and applying (2) directly to that ratio. The expression (13''') interests us because it leads to a useful approximation: the first term, which can be computed with simpler methods (6) and with data only from the current surveys, proves to be a good approximation for the entire expression. Thus the easily computed variance for a change in scores can be used to estimate the more difficult variance of a change in relatives.

The variance of $(R_2 - R_1)$ can be seen in (13') as consisting of the algebraic sum of the computing units $(e_{jh} - R_j e_{oh})$ from (10'') with the coefficient of +1 for j = 2 and -1 for j = 1. This ready result comes from the simple form of the variance of the relative, which consists of the sum for all strata of the squares of the computing units. Similarly, the variance of any simple weighted sum of the relatives $\sum_j^J w_j R_j$ takes the form

$$\mathrm{var}(\sum_j^J w_j R_j) \doteq \sum_h^H [\sum_j^J w_j (e_{jh} - R_j e_{ojh})]^2 \qquad (14)$$

These derivations are similar to those for formula (5). Squaring the brackets yields the proper terms of variances and covariances for all the j's. Our present interest is in two forms. One of these (with all $w_j = 1$) yields the <u>index</u> as the simple sum of the J relatives composing the index. Its variance is

$$\mathrm{var}(\sum_j^J R_j) \doteq \sum_h^H [\sum_j^J (e_{jh} - R_j e_{ojh})]^2 \qquad (15)$$

This seems like a good computing form. But for analytical purposes we also note, that, compared to the variance of a single relative (10''), the variance of an index is the sum of the variances and covariances of all (j = 1, 2...J) relatives:

$$\mathrm{var}(\sum_j^J R_j) \doteq \sum_h^H \sum_j^J (e_{jh} - R_j e_{ojh})^2 + 2\sum_h^H \sum_{j<k}^J (e_{jh} - R_j e_{ojh})(e_{kh} - R_k e_{okh}) \qquad (15')$$

$$= \sum_j^J \mathrm{var}(R_j) + 2\sum_{j<k}^J \mathrm{cov}(R_j, R_k) \qquad (15'')$$

Finally, we want the variance of the difference of two indexes $I_2 - I_1$; and this is (14) with all $w_{2j} = 1$ and all $w_{1j} = -1$:

$$\operatorname{var}\left(\sum_{j}^{J} R_{2j} - \sum_{j}^{J} R_{1j}\right) \doteq \sum_{h}^{H}\left[\sum_{j}^{J}\left\{(e_{2jh} - R_{2j}e_{ojh}) - (e_{1jh} - R_{1j}e_{ojh})\right\}\right]^2 \qquad (16)$$

$$= \sum_{h}^{H}\left[\sum_{j}^{J}\left\{(e_{2jh} - e_{1jh}) - (R_{2j} - R_{1j})e_{ojh}\right\}\right]^2 \qquad (16')$$

$$= \sum_{h}^{H}\left[\sum_{j}^{J}(e_{2jh} - e_{1jh})\right]^2 \qquad (16'')$$

$$+ \sum_{h}^{H}\left[\sum_{j}^{J}(R_{2j} - R_{1j})e_{ojh}\right]^2$$

$$- 2\sum_{h}^{H}\left[\sum_{j}^{J}(e_{2jh} - e_{1jh})\right]\cdot\left[\sum^{J}(R_{2j} - R_{1j})e_{ojh}\right]$$

The three terms correspond to the analogous terms in equations (13) for the variance of the difference of relatives, but comprise the summations for all (j = 1, 2,...J) relatives. The first term consists of the variances and covariances of the score differences $(r_{2j} - r_{1j})$ for the different relatives (j = 1, 2,...J) and including the factors r_{oj}^{-2} without sampling variation, i.e., as a constant for each j:

$$\sum_{h}^{H}\left[\sum_{j}^{J}(e_{2jh} - e_{1jh})\right]^2 = \sum_{j}^{J} r_{oj}^{-2}\operatorname{var}(r_{2j} - r_{1j})$$

$$+ 2\sum_{j<k}^{J} r_{oj}^{-1}r_{ok}^{-1}\operatorname{cov}[(r_{2j} - r_{1j}), (r_{2k} - r_{1k})] \qquad (16''')$$

We find (line 8, Table 3) that the covariances among the change of relatives are important and that they are in proportion to the covariances of the relatives, the second term of (15"). But we also find (line 7, Table 3) that this entire first term (16''') is an excellent approximation for all the three terms of (16") for the variance of index changes--just as the first term of (13''') was for the variance of changes of individual relatives. Thus the second and third terms of (16") can be neglected often--when the factors $(r_{2j} - r_{1j})/r_o$ are uniformly small enough to be neglected. They consist of sums of variances and covariances, generalizations for the J relatives of the second and third terms of (13''').

3. Standard Errors for Scores and Their Changes

The methods described in the preceding sections have been applied to a large body of empirical data. The chief aim was the computation of variances for long series of economic indicators as they have been used for about 12 years in the Surveys of Consumer Expectations [2]. Data from six surveys are represented, each of about n = 1370 interviews and two from each of three years: the base year 1956, and two current years, 1959 and 1960. Essentially the same methods of multi-stage probability area sampling methods were used in each survey to select dwellings with equal probabilities. In the dwellings families were identified (1.04 per dwelling) and a single interview was taken from head or wife alternately designated (or from head, if without a spouse).

The 1320 or so dwellings came from about 400 segments and these from 66 primary sampling areas, either single counties or Standard Metropolitan Areas--a widespread sample, with an average of 20 cases coming from an average of seven segments per area. Of the 66 areas 54 were selected with less than certainty and these contribute 27 strata, each containing two primary selections, to the computation of the variance. The other 12 are the largest metropolitan areas, included with certainty; here segments were the primary sampling units and these were paired into 18 computing strata. Altogether then the computation of the variance is based on 27 + 18 = 45 = H strata, with two primary selections in each [4].

The 15 variates represent answers obtained in essentially the same form over the years. All 15 variates, as they appear here, represent trinomials. For example on the first item the response "Better off" has a value 200, the negative response "Worse off" has a value of 0, and the neutral response has a value of 100; the neutrals responses are mostly "Same", with fewer "Uncertains" and "Don't Knows" and very few "Unascertained". The mean is 110 and this expresses results from about 30 per cent "Better off" against 20 per cent "Worse off", with 50 per cent in the middle group; the deviation from 100 expresses the difference between positive and negative percentages. All 15 items have similar positive directions, denoting optimism, improvement or expectations to buy. Of the eight items (1-8) of attitudes and expectations, seven average more than 100, showing more positive than negative responses.

A great difference of behavior separates the eight attitudinal variates (1-8) and the seven rare buying variates (9-15). This difference runs through the entire analysis. The attitudinal items have fairly large middle groups, comprising about one-third or one-half, and both sides are of moderate size. On the contrary, the seven buying items represent "J"-shaped distributions. The "certain or probable" buyers are rather rare and the "slightly" inclined or "uncertain" are even fewer, but nonbuyers are numerous. For example, a score of 14 might typically consist of these proportions: .06 buyers with scale value 200; .02 slightly inclined with scale value 100; and .92 nonbuyers with scale value 0. (Note that to estimate the actual proportion of buyers one should divide the "raw" scores and their standard deviations by two. But all coefficients and relatives remain unaffected thereby.) Thus the five variates (numbered 9, 12-15) with scores near 14 denote buying intentions of about seven per cent; variate 10 with scores of about 32 represents car buying by 16 per cent and variate 11 with scores around 42 represents home repairs by 21 per cent. There is little difference in the statistical properties of this kind of score and the simple use of the proportion of buyers, and the latter has been often preferred in the survey.

Table 1.--Standard Errors for Scores (r) and Score Changes ($r_2 - r_1$)

Item	Scores r	S.E.(r)	C_r = S.E.(r)/r x 100	S.E.($r_2 - r_1$)	Fluctuations in 17 Surveys (s.d.)
(1)	(2)	(3)	(4)	(5)	(6)
1. Evaluation of financial situation as compared with a year earlier	110	2.32 (.14)	2.11	2.95	6.26
2. Expected change in financial situation	128	1.73 (.10)	1.35	2.37	3.78
3. Business condition expected over the next 12 months	156	2.29 (.31)	1.47	2.91	12.93
4. Business condition expected for the next five years	122	2.35 (.35)	1.93	2.54	10.83
5. Good or bad time to buy large household goods	129	2.74 (.31)	2.12	3.10	8.50
6. Changing prices expected for next year is to the good or bad	92	2.37 (.33)	2.58	3.15	9.68
7. Evaluation of current business conditions compared to those a year ago	109	2.76 (.25)	2.53	3.72	27.37
8. Expected business condition a year from now as compared with the present	120	1.68 (.35)	1.40	2.26	11.96
9. Plan to buy house during the next year	14	1.39 (.27)	9.93	1.92	1.76
10. Intention to buy automobile during the next 12 months	33	1.87 (.20)	5.67	2.42	2.94
11. Evaluation of chances of home repair	42	2.72 (.32)	6.48	3.70	4.25
12. Evaluation of chances of buying refrigerator	14	1.48 (.13)	10.57	2.06	1.60
13. Evaluation of chances of buying T.V.	14	1.23 (.22)	8.78	1.61	1.29
14. Evaluation of chances of buying cooking range	12	1.17 (.17)	9.75	1.64	1.55
15. Evaluation of chances of buying washing machine	17	1.38 (.18)	8.12	1.92	1.69

The score in column 2 of Table 1, stands for the mean of the six "raw" scores obtained on the six surveys. In column 3 we present the mean of standard errors computed for the six studies; the standard deviations shown underneath in parentheses measure the variations of those six values. The standard errors are fairly uniform, mostly in the range of 1 1/2 to 2 1/2 points. Nevertheless differences do exist among them; some of these differences are appreciable--also reliably large compared to their own standard errors ($1/\sqrt{6} \doteq 0.4$ times their standard deviations).

Column 4 contains the coefficients of variations: the mean standard errors of column 3 divided by the mean scores of column 2. The uniformity we noted before in column 3 disappears here in column 4: the seven rare expectations (9-15) have much larger coefficients of variation compared to the attitudes (1-8). For the sources of this difference we may look to the coefficients of variation under unrestricted random sampling. For binomials this would be $\sqrt{(1-P)/Pn}$; for trinomials it can be written, with P_2, P_1, and P_0 denoting the proportions having scores of 2, 1, and 0 respectively, as:

$$\sqrt{\frac{4P_2 + P_1 - (2P_2 + P_1)^2}{(2P_2 + P_1)^2\, n}} = \sqrt{\frac{P_2 + P_0 - (P_2 - P_0)^2}{(1 + P_2 - P_0)^2\, n}} \qquad (17)$$

When $P_2 = P_0$ the coefficient of variation becomes $\sqrt{(P_2 + P_0)/n}$. The attitudes (1-8) approximate this condition; furthermore, they tend to be "humped in the middle" so that P_1 is of considerable size and $P_2 + P_0$ is in the neighborhood of 0.5. Under these conditions the coefficient of variation is $\sqrt{0.5/n} = 0.7/\sqrt{n}$ and for n = 1370 it comes to .019. However, for the rare items the situation is modified: when P_2 approaches 0 then the coefficient of variation approaches the binomial model $\sqrt{P_0/(1 - P_0)n}$. For example for the score of 14 with $P_2 = .06$, $P_1 = .02$, $P_0 = 0.92$, we get $\sqrt{(.98 - .86)^2/(1.06 - .92)^2 n} = 3.50/\sqrt{n}$. For $P_0 = .92$ the binomial is $3.39/\sqrt{n}$; and with $P_0 = .93$ (for a score of 14/2) it comes to $3.64/\sqrt{n}$. For n = 1370 the value of $3.50/\sqrt{n} \doteq .095$. These facts explain the source of the high coefficients of variation of the five rare items 9, 12-15 and items 10 and 11 fare only a little better.

We should add that when the trinomials are broken

into three separate binomials, these have large coefficients of variation. For example, the first item usually appears also in the surveys in terms of percentages of "Better off", "Worse off" and "Same plus Uncertain"; and these percentages have coefficients of about .04 to .07. While using the coefficients of variation insofar as they help we should do so with caution; they are affected by the arbitrariness of direction for vinomials and by the arbitrariness of both location and direction for the trinomials. But the coefficients also are important for understanding the variance of relatives and indexes.

Column 5 contains the standard errors of changes of scores $(r_2 - r_1)$ between two periods. Each entry is the mean of three computations obtained for the pairs of surveys from each of three years. We used formula (6) for the variance of the difference of two ratio means.

The standard errors for the differences in col 5 are somewhat higher than for the corresponding means (column 3), but not $\sqrt{2}$ greater, as they would be without positive correlation between the pairs of means. We computed the ratios of standard error of $(r_2 - r_1)/\sqrt{\text{var}(r_2) + \text{var}(r_1)}$; the computed ratios lay between 0.77 and 0.99 and average 0.916. Thus the positive correlation between the r_2 and r_1 reduces the standard error of $(r_2 - r_1)$, but leaves it greater than the variance of a single ratio. As a working average we may consider that the average ratio $0.916^2 = 0.84$ expresses the effect of an average correlation of $R = 0.16$--assuming $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$ and with $(\sigma^2 + \sigma^2 - 2R\sigma\sigma)/(\sigma^2 + \sigma^2) = (1-R) = 0.84$ These results resemble hundreds of similar computations we have made on many other surveys, but still unpublished.

To measure the fluctuations of these "raw" scores between the periodic surveys we used the scores for 17 surveys available in the years 1955-1962; the standard deviations between the 17 survey scores appear in column 6. Some of the fluctuation is due to sampling variations and it is interesting to compare the extent of fluctuations of scores between surveys with their standard errors. The attitudes (1-8) all had fluctuations much greater than the standard error, though variate 2 fluctuates less than the others and variate 7 much more. But the seven buying expectations (9-15) had fluctuations only about as great as their standard errors. These fluctuations could not individually contribute much new information; the usual fluctuations are about equal to the standard errors of the differences. But to conclude that the measures are useless would be too hasty. First, the variates can detect the larger fluctuations, which perhaps are more important though less frequent. We should add here that the range of fluctuation of all 15 items was on the average (with some variation) about 3.5 as great as the standard deviations presented here. (These results for the range/s.d. are about what one could expect for a sample of 17 from normal distributions.)

Second, if we would regard the fluctuations as "random", then the values of column 6 should be multiplied by $\sqrt{2} = 1.4$ before comparing them with the entries of column 5; on this fairer view the variates appear to contain some information.

Because the sample was clustered (and multi-stage) the standard errors actually computed in conformity with the sample design were greater than would appear from formulas proper only for unrestricted random sampling. This ratio of the actual standard errors to the unrestricted random model we denote as $\sqrt{\text{Deff}}$; its values were computed and their means assessed for each of the 15 variates as 1.21, 1.16, 1.42, 1.24, 1.33, 1.07, 1.46, 1.28, 1.08, 0.99, 1.32, 1.14, 0.97, 0.97, 0.96. The average of the 15 values is 1.17, with some variation among them; this represents an increase of $1.17^2 = 1.37$ in the variance. Its chief probable sources are the clusters of about three interviews in somewhat homogeneous segments; clusters of about 20 in less homogeneous counties; and clusters of about 10 interviews per interviewer [6]. These magnitudes are in accord with dozens of other variates computed for these surveys; also with hundreds of other computations on many other SRC surveys of national household samples.

We had also computed average values of $\sqrt{\text{Deff}}$ for the score changes, by taking the the actual standard errors of $(r_2 - r_1)$ as the ratio of the standard error of two independent unrestricted random samples of the same numbers of interviews; for the 15 items they were 1.07, 1.11, 1.28, 0.96, 1.08, 0.99, 1.36, 1.15, 1.04, 0.90, 1.25, 1.10, 0.88, 0.95, 0.95. We find that effects of clustering remain (averaging 1.07 for the standard error and $1.07^2 = 1.15$ for the variance) but are not as great as for the single ratio means. The reduction of the effect is due to the positive correlation between the two means. Again, we have found these relationships for dozens of other variates on these surveys and for hundreds on many other surveys; though still not published separately. The correlations come from using the same counties and some but not all of the same blocks and segments; using the same interviewers also reduces their effect if present in the separate means [6].

In column 3 we not only gave the means of the six replicate computations of standard errors, but also in parentheses the standard deviations between the six values for each. We took advantage of the information from the repeated computations performed on what are essentially six replications to obtain estimates of the coefficient of variation for the standard error. The computation of the standard errors should be subject theoretically and approximately to a coefficient of variation of about $1/\sqrt{2 \times 45} = .106$, because the 45 pairs of comparisons result in about 45 degrees of freedom. We should actually expect somewhat more, because of probable skewness of $\beta_2 > 3$ and because the 90 primary selections vary somewhat in size. The 15 computed values vary from .06 to .20 and they average .128. We should expect this to be an overestimate because of variations between

surveys in the measurements and in the sample designs. A similar coefficient of variation applied to the values of $\sqrt{Deff}$ averages .120, also varying from .06 to .20; the difference is fortuitous though in the expected direction; we expect $\sqrt{Deff}$ to vary somewhat less than the sampling error because here the variations in sample size and in the level of the variates between surveys are largely eliminated. But we can use .12 as a working estimate of the coefficient of variation of the standard errors.

Some readers may be interested in the technical bias of the ratio mean r. These were computed and all found to be negligible: much less than one per cent of the values of the standard errors and appeared, when tested, to vary haphazardly around zero. This was expected in light of previous research [5] and in light of the computed coefficient of variation of about .04 for the sample size x.

4. Standard Errors for Relatives and Their Changes

The results about relatives in this section, also about their sums, the indexes in the next section, were computed for four surveys, two in 1959 and two in 1960. The two surveys of 1956 combined serve as the base period. The standard errors of relatives were computed, using (10") for each of the four surveys. The means of the four standard errors appear in column 1 of Table 2. The chief lesson again concerns the striking contrast between the eight attitudinal items (1-8) on the one hand, and the seven rare items (9-15) on the other. The standard errors for the former range from about 1.6 to about 3.1; for the latter they range from 6.0 to over 13.

This pattern follows the behavior of the coefficient of variation of the simple scores (col 4 of Table 1). The standard errors of the relatives are high for those items that have large coefficients of variation of the score. Because the relatives have values near 100 their coefficients of variations are much like their standard errors times the factor 100.

In column 2 of Table 2 we present standard errors for the changes $(R_2 - R_1)$ of two relatives. Each entry is the mean of six standard errors computed for the differences between the six possible pairs of the four periods investigated. We note that the standard errors for items 1-8, are again much smaller than for the seven rare items. We should also note the remarkable fact that these entries are only a little greater than the corresponding entries in column 1 for the standard errors of separate relatives. The variances of the differences are reduced by the strong correlations between the pairs of relatives R_2 and R_1. These effects may be seen in the entries of column 3, each showing the ratio of the standard error of difference $(R_2 - R_1)$ to the square root of the sum of the two variances for R_2 and R_1; each entry is the mean of the ratios for the six samples involved. The ratios are well less than 1, which would denote the absence of correlation; their average is 0.800, which denotes a variance ratio of 0.64, the effect of a correlation of 0.36 on the average.

We also obtained most reassuring confirmations for the conjecture that the variance of the change in the relatives $(R_2 - R_1)$ may usually be computed more simply in terms of the variance of change in the scores $(r_2 - r_1)$, taking advantage of the approximation that $r_o^{-2}\text{var}(r_2 - r_1) \doteq \text{var}\left(\frac{r_2 - r_1}{r_o}\right) = \text{var}(R_2 - R_1)$. We computed the ratios $[\text{var}(R_2 - R_1)/r_o^{-2}\text{var}(r_2 - r_1)]^{1/2}$ and averaged these over the six possible pairs for each of the items. All 15 ratios of the standard errors were found within .014 of 1.00; there was little variation and the mean of the 15 was 1.00. Thus we can safely use this approximation when $(r_2 - r_1)/r_o$ is moder-

Table 2.--Standard Errors of Relatives $(R_1 = r_1/r_o)$ and Changes of Relatives $(R_2 - R_1)$

Item	S.E.(R)	S.E.$(R_2 - R_1)$	Fluctuations in 17 Surveys (s.d.)	$\frac{\text{S.E.}(R_2 - R_1)}{\sqrt{[\text{Var}(R_2) + \text{Var}(R_1)]}}$	$\frac{C_R}{\sqrt{C_{r_o}^2 + C_r^2}}$
	(1)	(2)	(3)	(4)	(5)
1	2.30	2.82	5.40	.87	.88
2	1.65	1.78	2.97	.76	.96
3	1.59	1.86	7.89	.82	.94
4	1.79	2.01	7.94	.79	.88
5	2.42	2.23	6.29	.65	.95
6	2.89	3.62	7.45	.89	.96
7	3.11	3.66	24.64	.84	.97
8	1.90	2.21	17.26	.81	.90
9	11.93	14.24	12.69	.83	.95
10	6.02	6.83	8.52	.80	.90
11	7.30	8.21	9.93	.79	.95
12	13.37	14.71	11.18	.78	.98
13	10.00	11.29	8.57	.79	.98
14	13.14	14.95	14.17	.80	1.00
15	11.44	12.40	11.10	.77	.98
Mean				.80	.944

ately small. This holds usually, especially when the standard error is needed. Then we may simply compute the variance of the score change instead of the more difficult variance of the change in relatives. We can note also that decreasing the variance of r_o does not decrease the variance of the changes in relatives.

Searching also for a simpler approximation for the variance of the relative, we wondered how important the correlation $\rho_{r_o r_o}$ was in $C_R^2 \doteq C_r^2 + C_{r_o} - 2\rho_{r_o r_o} C_{r r_o}$. We found that the ratios $[C_R^2/(C_r^2 + C_{r_o})]^{1/2}$ varied only between 0.88 and 1.00, averaging 0.944. It seems then that using the convenient values of $(C_{r_o}^2 + C_r^2)$ results in overestimating C_R, the coefficient of variation of the relative, on the average by about 5 1/2 per cent (the variance by 11 per cent) and it also appears to be a fairly stable substitute. (This agrees with the slightly greater correlation of 0.16 found between successive surveys.) Hence, knowing C_{r_o} for the base, by computing the relatively simple estimates C_r we can obtain fairly good estimates of the more difficult C_R. We also note that by decreasing the variance of r_o, we decrease the variance of every relative based on it. In our case r_o was based on two samples and $C_{r_o}^2$ is about 0.5 C_r^2; the small correlation (0.16) between the two samples is counteracted by the circumstance that the r_o's in the base period seem greater than the average r since. Thus approximately $C_{r_o}^2 + C_r^2 \doteq 1.5\, C_r^2 \doteq 3\, C_{r_o}^2$.

Thus $\mathrm{Var}(R) = C_R^2 \doteq 0.9(C_r^2 + C_{r_o}^2 \doteq 0.9(1.5C_r^2)$ $= 1.35C_r^2$. But we also found that $\mathrm{var}(R_2 - R_1) \doteq r_o^{-2}\mathrm{var}(r_2 - r_1) \doteq 2\mathrm{var}(r)(1 - 0.16)$, because $\rho_{r_2 r_1} = 0.16$ on the average. Thus assuming $r/r_o = 1$ on the average we should expect $\mathrm{Var}(R_2 - R_1)/[\mathrm{Var}(R_2) + \mathrm{Var}(R_1) \doteq 0.84/1.35$ $= 0.62 = 0.79^2$. This explains the sources of the average value of 0.80 found in column 4. The variance of the base r_o contributes to the denominator but not to the numerator.

We took advantage of the four replicated computations of the standard error of each relative to estimate their variability. The coefficients of variations of the standard errors averaged 11.8, resembling closely the results for the scores.

5. Standard Errors for Indexes and Their Changes

Table 3 presents results regarding five indexes in separate columns, each index being the sum of several relatives. The first index is the sum of the relatives for the six expectations (1-6); the second index contains only items 9 plus 10 (home and car buying); the third is the sum of these eight items. The fourth is the sum of the four intentions to buy appliances (12-15); the fifth adds items 9 and 10 to them; these two indexes have not been actually used but only contemplated. Conforming to usage, instead of sums the means are computed, so that the indexes take the form $I = \frac{1}{J}\sum^{J} R_j$ and fluctuate around 100 as the relatives do.

Table 3.--Factors Relating to the Indexes ($I = \frac{1}{J}\sum^{J} R_j$) and Index Changes: $I_2 - I_1 = \frac{1}{J}\sum^{J} R_{2j} - \frac{1}{J}\sum^{J} R_{1j} = \frac{1}{J}\sum^{J}(R_{2j} - R_{1j})$

	Items in Index				
	1-6	9,10	1-6, 9,10	12-15	9,10 12-15
1. Fluctuations (s.d.) of I in 17 Surveys	4.16	8.33	4.63	7.62	5.01
2. Standard Errors of the I	1.165	7.171	2.110	7.714	5.505
3. Effects of Correlations on Var(I)	1.754	1.129	1.372	1.626	1.796
4. Mean Correlations among the R_j	0.161	0.174	0.119	0.213	0.152
5. Standard Errors of Change (I_2-I_1)	1.295	8.199	2.381	8.922	6.217
6. $\mathrm{Var}(I_2 - I_1)/$ $[\mathrm{Var}(I_2) + \mathrm{Var}(I_1)]$	0.626	0.643	0.628	0.661	0.635
7. Ratio of Var to Simple Approximation	0.992	1.035	1.029	1.007	0.996
8. Effects of Correlations on $\mathrm{Var}(I_2 - I_1)$	1.657	1.080	1.269	1.747	1.816

The standard errors are on line 2, each the mean for the four periods computed. The results here conform to the results presented in Table 2 for the separate relatives. The index for items 1-6 has a very low standard error; that for items 9 plus 10 is much higher--so much higher that adding them to the first six items increases substantially their standard error. Indexes based on the highly variable items 12-15 would also have high standard errors.

On line 3 we investigate the effect of increased variances due to the correlations between the relatives composing an index, computed as $\mathrm{Var}(I)/J^{-2}\sum^{J}\mathrm{Var}(R_j)$. In the first column for example we have the variance of $I = \frac{1}{6}\sum_{j=1}^{6} R_j$ for items 1-6 divided by $\frac{1}{36}\sum_{j}^{6}\mathrm{Var}(R_j)$ and find this ratio to be 1.754 (actually the mean of such ratios computed for the four periods). In the absence of correlation between the R's, this ratio would be unity and the index would have a variance 1/6 as large a single relative on the average. Actually since the ratio is 1.754 the variance of the index is smaller in the proportion 1.764/6 = 1/3.42; the

correlations reduce the information in the six items from 6 to the equivalent of 3.42 items.

These considerations led us to compute an approximate and synthetic correlation coefficient as:

$$\frac{\text{Variance}(\sum R_j) - \sum \text{Var}(R_j)}{[\sum \sqrt{\text{Var}(R_j)}]^2 - \sum \text{Var}(R_j)} = \frac{\sum\sigma_j^2 + \sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j - \sum\sigma_j^2}{\sum\sigma_j^2 + \sum_{i\neq j}\sigma_i\sigma_j - \sum\sigma_j^2} = \bar{\rho}_{ij} \qquad (18)$$

The results, the means of four computations, appear in row 4. The last step is justified only insofar as the variances are about equal. This condition obtains fairly well for the first index; also for the fourth and fifth indexes; it clearly fails for the third index which mixes the low variance items 1-6 with the high variance items 9 and 10. The "decay of information" (the increase of the variance) due to positive correlations can be put in the form $1 + \rho(J-1)$; and the mean of J correlated items will have a variance $\frac{1}{J} + \rho\frac{J-1}{J}$. It is encouraging to note that the average correlation between the six attitudes (1-6)--also between their changes--is only about 0.16. This should overcome any suspicion that the several questions might merely yield the individual's general feeling of optimism or pessimism. It is also interesting that the buying behaviors (12-15) are positively correlated with an average of 0.21; this may be due to either general financial ability or the positions in the life cycles of individual consumers. These points could be better studied by noting the correlations for individual consumers.

Rows 5-8 of Table 5 contain results about changes in the index $(I_2 - I_1)$ between two periods. Each of the entries again represents the mean of six computations for the six possible differences between four periods. The outstanding fact to note in this table is how well preserved are the several useful relationships we had noted before, either for the single index or for the changes $(R_2 - R_1)$ of single relatives.

Note on line 5 that the standard errors of differences are only slightly greater than the standard errors of individual indexes on line 2. This relationship is similar to that found between the standard errors for the relative R and for the changes $(R_2 - R_1)$. Again this phenomenon is due to the high correlation between I_2 and I_1. Its effect is measured by the ratio of the $\text{Var}(I_2 - I_1)$ to the sum of the variances of I_2 and I_1; note on line 6 that these come close to $0.64 = 0.80^2$, the value we found as the average in column 4 of Table 2.

On line 7 we find again, as we found for the difference $(R_2 - R_1)$ of relatives, that the variance of $(I_2 - I_1)$ can be approximated very well with the variance of the mean of score changes over the base r_{oj}; the entries give the ratios of the variance of $(I_2 - I_1)$ to the variance of $[\frac{1}{J}\sum(r_{2j} - r_{1j})/r_{oj}]$. Again these approximate standard errors would be close enough, within 1 1/2 per cent. The values of r_{oj} are known in advance; used as weights in (18) this approximation leads to easier computations for the more complicated variance of $(I_2 - I_1)$. It can also lead for easy approximations for the variance of the indexes I, insofar as the factor 0.64 of line 6 can be considered reliable.

Finally, on line 8 we investigate the increase of the variance due to correlations among the J different changes of $(R_{2j} - R_{1j})$ the sum of which is the index change $(I_2 - I_1)$, computed as $\text{Var}(I_2 - I_1)/J^{-2}\sum \text{Var}(R_{2j} - R_{1j})$. The increases seen here for the variances of index changes are similar to the increases noted on line 3 for the separate indexes.

6. The Statistical Bias of the Relative

We expect the ratio mean $R = r/r_o$ of the relative to have a ratio bias. We expect it because the bias is known (see, e.g., formula 4 in [5]) to have this relationship to the standard error: $\text{Bias}(R)/\text{S.E.}(R) = \rho_{Rr_o} C_{r_o}$. Because we expect the correlation ρ_{Rr_o} to be negative we should get a positive bias. But we can also expect its ratio to the standard error to be less than C_{r_o}, the coefficient of variation of the base score r_o.

Knowing that the statistical bias exists, we want to know how large it is. For computing estimates of the bias we used the relationship:

$$\text{Bias}(R)/R \doteq C_{r_o}^2 - C_{rr_o} \doteq (1/2)(C_R^2 - C_r^2 + C_{r_o}^2) \qquad (19)$$

This may be obtained from formulas 2 and 3 in [5] or 4.11.1 and 4.14.2 in [1, Vol II]. The data show clearly the existence of the positive bias, for each of the computed values of the relative bias had a mean about 3 to 8 times as large as the standard deviation between the four estimates. But all the biases were small; for the eight attitude items from .01 to .03 of one per cent, and for the seven rare buying items from 0.1 to 0.6 of one per cent.

A more relevent comparison is the ratio of the relative bias to the coefficient of variation; $[\text{Bias}(R)/(R)]/C_R = \text{Bias}(R)/\text{S.E.}(R)$. These values of the bias ratio were: 0.80, 0.57, 0.49, 1.03, 0.56, 1.06, 0.51, 5.00, 1.92, 2.08, 4.80, 3.89, 3.65, 3.90. Note the great difference again between the first eight and the last seven items. But even the maximum bias ratio is only 5 per

cent; even this is negligible since it increases the mean square error by only $(1 + .05^2) = 1.0025$.

The bias ratio equals $-\rho_{Rr_o} C_{r_o}$ and based on relationships and estimates already discussed we expect on the average:

$$\rho_{Rr_o} \doteq \rho_{rr_o} \frac{C_r}{C_R} - \frac{C_{r_o}}{C_R} \doteq .109 \frac{1}{\sqrt{.9 \times 1.5}}$$

$$- \frac{1}{\sqrt{.9 \times 3}} = .086 - .61 = -.52$$

The computed values ranged from 0.37 to 0.61, averaging 0.52; these were obtained by dividing the bias ratio by C_{r_o}. The values of C_{r_o} were 1.68, 1.04, 1.11, 1.32, 1.73, 1.45, 1.88, 1.10, 8.32, 4.12, 4.45, 8.01, 6.77, 6.61, 6.38. (These values tend to be a little greater than 0.707 times the values of C_r in column 4 of Table 1)

The regular relation of the bias ratio to C_{r_o} has several consequences: (1) The bias ratio can be kept low by keeping C_{r_o} low; note the low bias ratios and the low C_{r_o} values for the first eight items. (2) The two biases tend to cancel from the difference $(R_2 - R_1)$ of two relatives. (3) When summed in indexes the bias ratios tend to grow, because the biases sum up, whereas the standard errors don't increase as fast. The worst case we can make out is for the six-item index in column 5 of Table 3; the ratio of relative bias to the coefficient of variation is $0.438/5.505 = .080$ and this would increase the mean square error by the factor $1 + .080^2 = 1.0064$. (4) The predictability of the relation permits adjustment when it becomes desirable.

7. Some Conclusions

This was not chiefly a theoretical exercise with an empirical illustration; rather we tackled an important practical problem with whatever tools we could find, invent and improvise.

We must be more circumspect and specify that we studied only sources of sampling variations, leaving to others the problems of validity and the predictive value of the indicators including the problems of criteria for testing them.

But knowledge of the sampling variations and its sources provides the necessary base for further investigations, and the results also have important direct implications. As suspected the rare buying intentions have large sampling errors and are useful only for detecting large changes. How could one improve their sensitivity? Perhaps with research the scope of the interview could be enlarged and the sensitivity increased in the middle of the scale of intentions. Most likely however only a drastic increase in sample size will prove useful. This perhaps could be achieved with some special auxiliary methods.

Most important are the results showing the excellent performances of all eight attitude items. They were obtained with precisions entirely adequate for their purpose and remarkable in light of the small sample sizes. Their performances are particularly good for measuring changes. Furthermore, the cumulative information they appear to provide when summed into an index is particularly gratifying.

REFERENCES

[1] Hansen, M.H., Hurwitz, W.N. and Madow, W.G., Sample Survey Methods and Theory, New York: Wiley and Sons, (1953).

[2] Katona, G. and Mueller, E., Consumer Expectations, Ann Arbor: Institute for Social Research, (1957).

[3] Keyfitz, N., "Estimates of Sampling Variance Where Two Units Are Selected from Each Stratum," Journal of the American Statistical Association, 52 (1957), 503-510.

[4] Kish, L. and Hess, I., "On Variances of Ratios and Their Differences in Multi-Stage Samples," Journal of the American Statistical Association, 54 (1959), 381-392.

[5] Kish, L., Namboodiri, N.K. and Pillai, R.K., "The Ratio Bias in Surveys," Journal of the American Statistical Association, 57 (December 1962).

[6] Kish, L., "Studies of Interviewer Variance for Attitudinal Variables," Journal of the American Statistical Association, 57 (1962), 92-115.

[7] Rao, C.R., Advanced Statistical Methods in Biometric Research, New York: Wiley and Sons (1952).

[8] Yates, Frank, Sampling Methods for Censuses and Surveys, Third Edition, London: Charles Griffin and Company Ltd., (1960).

A TEST OF HOMOGENEITY FOR A STRATIFIED SAMPLE

Tomas Garza-Hernandez, C-E-I-R de Mexico, and
Philip J. McCarthy, Cornell University

A common problem in the analysis of data in the social sciences is that of testing homogeneity on distributions of qualitative variables. This problem arises when a population has been divided into several so called "domains", and at the same time a qualitative variable is defined on each domain. In this situation it is often required to compare the distribution of the qualitative variable across these groups, that is, to test whether the proportion of elements belonging to each category of the qualitative variable is the same in each of the domains.

In general, the distribution of the population elements among the several domains is not known beforehand. This fact introduces a complication in the analysis, since it is not possible then to design a sampling procedure which will yield a specified number of elements from a given domain. Moreover, it sometimes happens in practice that a prior stratification, unrelated to the domains of study, has been used in selecting the sample, thus introducing more difficulties in the solution of the problem, which of course remains unchanged, since the stratification is not relevant to the purposes of the investigation.

An example of the situation described above is that of a survey conducted in 1952 in a Canadian Maritime Province, whose specific aims were to assertain the incidence of psychiatric disorders within different subgroups of the population under study. This survey was conducted in the following manner (1):

1. A county was divided into three strata, corresponding distinct geographical and social areas. From each stratum according to some preassigned sampling rates, a sample of households was selected, and either the male or female in the household was interviewed.

2. The set of domains of study corresponded to categories of an index called "Occupational Disadvantage", roughly a measure of average well-being in various occupational levels. For example, one domain corresponds to owner, salaried and professional occupations, and another to self-employed workers in agriculture and fishing.

3. The qualitative variable is an overall judgment made by the project psychiatrists: a person belongs to category 1 if he is a well person, to category 2 if he would almost qualify for psychiatric attention or therapy, and to category 3 if the diagnosis is doubtful.

In this example, the hypothesis to be tested is that the proportion of persons within each category of "psychiatric status" is the same for both categories of "occupational disadvantage", the strata to play no role in the analysis.

A first approach to the solution might be through the application of a χ^2-test to the separate strata, and then combining the results for all the strata taken together, as described, e.g., by Kendall (2). But perhaps an overall test is required, relating to the entire population, regardless of the stratification, among other reasons because a χ^2-test would be meaningless whenever no observations are obtained from a given domain, and this may easily happen in an actual situation.

Let N_i, $i=1,2,\ldots,L$, denote the number of elements in the i-th stratum, and within this particular stratum, let $N_{ia}, N_{ib}, \ldots, N_{id}$ denote the number of elements in domains a, b, ..., d, respectively.

Within each stratum, say the i-th, and within domain a, say, in this stratum, let $P_{ia(1)}, P_{ia(2)}, \ldots, P_{ia(k)}$ denote the proportions of elements pertaining respectively, to categories 1, 2, .., k.

Then, in the whole population, we are interested in comparing the proportions $P_{a(j)}$, $j=1,2,\ldots,k$, where

$$P_{a(j)} = \frac{\sum N_i \pi_{ia} P_{ia(j)}}{\sum N_i \pi_{ia}}$$

and where we denote N_{ia}/N_i by π_{ia}, for domain a, and similarly for other domains.

We now want to consider the following problem: To test the hypothesis H: $P_{a(j)}=P_{b(j)}=\ldots=P_{d(j)}$, for each j=1, 2, ..., k, on the basis of a sample drawn at random from the separate strata.

In tabular form, we have the following situation:

$P_{a(1)}$	$P_{a(2)}$	----	$P_{a(k)}$
$P_{b(1)}$	$P_{b(2)}$	----	$P_{b(k)}$
$P_{d(1)}$	$P_{d)2)}$	----	$P_{d(k)}$

and it is desired to test whether the quantities appearing in a given column are the same no matter what the actual value is, that is, if the distribution of elements among the categories is the same for all domains.

Noting that the sum across columns equals 1 for any one row, we may leave the last column out of the analysis, and we then state our hypothesis as follows:

$$H: \quad P_{a(j)}=P_{b(j)}= \ldots = P_{d(j)}$$

simultaneously for $j=1,2, \ldots, k-1$.

a. Estimates of the parameters: their variances and covariances.

In accordance with the usual approach of finite population methodology, the estimates that we propose for the above parameters are simply given by

$$p_{a(j)}=\hat{P}_{a(j)}= \frac{\sum N_i \frac{n_{ia}}{n_i} p_{ia(j)}}{\sum N_i \frac{n_{ia}}{n_i}},$$

$j = 1,2,\ldots,k-1.$

where n_i is the size of the sample drawn at random from the i-th stratum, n_{ia} is the number of elements of the n_i which fall in domain a, and $p_{ia(j)}$ is the proportion of elements in domain a, stratum i, that fall in the j-th category.

The variance of this estimate follows, after some manipulations:

$$\mathrm{Var}(p_{a(j)}) \doteq \frac{1}{N_a^2} \sum \frac{N_i^2 (N_i-n_i)}{n_i (N_i-1)} \pi_{ia} P_{ia(j)}(1-P_{ia(j)}) (1-\pi_{ia}) \pi_{ia} (P_{ia(j)}-P_{a(j)})^2 \quad \text{where } N_a=\sum N_{ia}$$

This formula agrees with Hartley's (3), p. 15, formula (33), and although Professor Hartley remarks that it "appears to be restricted to a proportional allocation of the sample to strata", we have derived it with no assumptions on the sample allocation.

There are two kinds of covariances between our estimates. The first one arises since the sample of size n_i is composed of $n_{ia}, n_{ib}, \ldots, n_{id}$, all these adding up to n_i, thus giving correlations across domains. The second one comes from the fact that in a given stratum we have $\sum P_{ia(j)}=1$, and hence there are correlations between estimates across categories.

It can be shown, then, that

$$\mathrm{Cov}(P_{a(j)}, p_{b(j)}) \doteq -\frac{1}{N_a N_b} \sum \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i - 1} \pi_{ia} \pi_{ib} (P_{ia(j)}-P_{a(j)})(P_{ib(j)}-P_{b(j)}),$$

and

$$\mathrm{Cov}(P_{a(j)}, P_{a(\ell)}) \doteq \frac{1}{N_a^2} \left\{ -\sum N_i^2 \frac{N_i - n_i}{N_i - 1} \frac{\pi_{ia} P_{ia(j)} P_{ia(\ell)}}{n_i} + \sum N_i^2 \frac{N_i-n_i}{N_i-1} \frac{\pi_{ia}(1-\pi_{ia})}{n_i} (P_{ia(j)}-P_{a(j)})(P_{ia(\ell)} - P_{a(\ell)}) \right.$$

b. Derivation of the test procedure.

A test procedure for the stated hypothesis should be based on the behavior of the estimates in the sampling process. This behavior is described by specifying the joint probability distribution of the estimates. No attempt will be made to derive the exact form of this distribution; instead we shall make an assumption concerning the joint distribution of the d(k-1) random variables $p_{a(j)}, \ldots, p_{d(j)}$; $j=1,2,\ldots,k-1$.

For the sake of simplicity, let us rewrite the parameters $P_{a(1)}, P_{b(1)}, \ldots, P_{d(k-1)}$ as $P_1, P_2, \ldots, P_{d(k-1)}$, the same change is notation holding for the estimates of the parameters. In the new notation our hypothesis is now

$H: P_j=P_{j+1}=\ldots=P_{j+d-1}$, for each $j=1, d+1, 2d+1, \ldots, (k-2)d+1$.

Making a straightforward generalization of the often used approximation to the distribution of a proportion, or of a difference of proportions, we propose the following assumption: Let P denote a column vector whose components are $P_1, P_2, \ldots, P_{(k-1)d}$, and p denote a column vector whose components are $p_1, p_2, \ldots, p_{(k-1)d}$. Let A denote the $(k-1)d \times (k-1)d$ matrix whose element in the (i, j)-th place is the covariance between p_i and p_j. Then, generalizing the result for the univariate case, let the joint density of the random variables $p_1, p_2, \ldots, p_{(k-1)d}$ be given by

$$\varphi(p,P) = c \cdot \exp\left\{-\frac{1}{2}(p-P)' A^{-1} (p-P)\right\} \quad \ldots\ldots(1)$$

a multivariate normal density, where c is a constant, and A^{-1} is the inverse matrix of A.

A likelihood-ratio criterion will now be

derived to test H. The maximum of $\varphi(p,P)$ over the whole parameter space is clearly seen to be attained for p=P, and let it be denoted by φ_Ω

In order to find the maximum of $\varphi(p,P)$ subject to the condition that H: $P_j = P_{j+1} = \ldots = P_{j+d-1}$, for j=1, d+1, ..., (k-2) d+1, we proceed as follows: Since the problem is equivalent to finding the maximum of the exponent in (1), we shall use the method of Lagrange multipliers, trying to express the restrictions under which the minimum is to be attained in vector form.

Let $S_1, S_2, \ldots, S_{(d-1)(k-1)}$ be a set of vectors, where S_1 has 1 as its first component, -1 as its second component and zeros all the way down. S_r is constructed from $S_{(r-1)}$ by shifting all its elements one place downwards, and replacing the first one by zero. Then, the product $S_1'P=0$ expresses the fact that $P_1=P_2$. Also, $S_2'P=0$ means $P_2=P_3$, and so on until we express $P_{d-1}=P_d$

The minimization procedure is then carried out by the usual method:

Let
$D=(p-P)'A^{-1}(p-P)+m_1 S_1'P+m_2 S_2'P+\ldots+ m_{(k-1)(d-1)} S'_{(k-1)(d-1)}P$, where the m's are constants.

Taking the partial derivative of D with respect to P, and setting it equal to zero, we have the set of (k-1)(d-1) +1 simultaneous equations:

$$\frac{\partial D}{\partial P} = 0$$

$$m_1 S_1'P = 0$$

$$m_{(k-1)(d-1)} S'_{(k-1)(d-1)} P=0$$

The solution of this set of equations gives the value of P which maximizes (1) subject to H. Let this maximum be φ_ω.

Then, the test criterion is as follows: If $\lambda = \varphi_\omega / \varphi_\Omega$ is greater than or equal to λ_0, reject H, where λ_0 is chosen according to the level of significance desired, and using the fact that $-2 \log\lambda$ follows a χ^2-distribution with d(k-2) degrees of freedom.

REFERENCES

1. Hughes, C.C., et al. (1960). People of Cove and Woodlot. New York: Basic Books, Inc.

2. Kendall, M. G. (1945-46). The Advanced Theory of Statistics. London: C. Griffin.

3. Hartley, H. O. (1959). Analytic Studies of Survey Data. Reprint Series #63. Statistical Laboratory, Iowa State University, Ames, Iowa.

MULTIPLE FRAME SURVEYS

H. O. Hartley

1. Concepts.

The situations encompassed by the term 'multiple frame surveys' may be described as follows: In sample survey methodology one often finds that a frame known to cover approximately all units in the population is one in which sampling is costly while other frames (e.g. special lists of units) are available for cheaper sampling methods. However, the latter usually only cover an unknown or only approximately known fraction of the population. The paper develops a general methodology of utilizing any number of such frames without requiring any prior knowledge as to the extent of their mutual overlap.

2. Some history. (For more history see Sect. 9.)

The technique of multiple frame surveys has been used in the past occasionally and under special circumstances. For example, the 1960 Survey of Agriculture of the Bureau of the Census uses two frames, namely (A) a frame based on the conventional 'area sampling' approach; (B) a frame of farms conceptually and operationally 'associated' with the A-1 listings of the last (1959) Census of Agriculture.* Earlier, the Statistical Laboratory at Iowa State University had used a two frame approach in a small study of 'Effects of Industrialization on Farming' which was carried out for the Department of Economics and Sociology of Iowa State University. Here the two frames consisted of

(A) The customary rural area frame for sampling farm operators.

(B) Employees of the Clinton Motor Company who are also farm operators.

The combined use of these frames proved a successful combination for simulating screening and providing coverage.* There are, no doubt, other occasions when several frames have been used in the past. Although such isolated instances of the use of the method provides valuable experience, it is felt that no systematic methodology for the analysis of such surveys is available.

3. Definitions for two frames.

To fix the ideas, consider first two frames A and B and assume that a sample has been drawn from each frame. The sample designs may be entirely different in the two frames but the following assumptions are made:

(i) Every unit in the population of interest belongs to at least one of the frames.

(ii) It is possible to record for each sampled unit whether or not it belongs to the other frame.

This means we can divide the units of the sample into three $(2^2 - 1)$ domains.

Domain (a) The unit belongs to Frame A only

Domain (b) The unit belongs to Frame B only

Domain (ab) The unit belongs to both frames

*Information from unpublished memoranda.

The units in the population are also conceptually divided into the above domains. We now distinguish four different situations concerning our state of knowledge of the total number of units in the frames and in the domains and of our ability to allocate prescribed sample sizes to the domains.

Schedule 1. Principle situations in multiple frame surveys*

Case No.	Knowledge of population numbers in domains and frames
1	All domain sizes N_a, N_b, N_{ab}, etc. are known
2	All domain sizes N_a, N_b, N_{ab}, etc. are known
3	Domain sizes are not known, but frame sizes are known
4	Neither domain sizes nor frame sizes are known, but the relative magnitude of the frames is known

Case No.	Possibility of fixed sample allocations to domains and frames	Nature of domains
1	It is feasible to allocate prescribed sample sizes to domains	Domains = strata
2	Prescribed sample sizes can only be allocated to frames	Domain = post-strata
3	Prescribed sample sizes can only be allocated to frames	Domains = domains proper
4	Prescribed sample sizes can only be allocated to frames	Domains = domains in populations of unknown size

*For explanation of symbols see Schedule 2.

We now introduce a convenient notation for the two-frame surveys.

Schedule 2. Notation for two frame designs and estimates

	Frame		Domain		
	A	B	a	b	ab
Population number	N_A	N_B	N_a	N_b	N_{ab}
Sample number*	n_A	n_B	n_a	n_b	n'_{ab}, n''_{ab}
Population total	Y_A	Y_B	Y_a	Y_b	Y_{ab}
Population mean	$\bar{Y}_A$	$\bar{Y}_B$	$\bar{Y}_a$	$\bar{Y}_b$	$\bar{Y}_{ab}$
Sample total*	y_A	y_B	y_a	y_b	y'_{ab}, y''_{ab}
Sample mean*	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_a$	$\bar{y}_b$	$\bar{y}'_{ab}$, $\bar{y}''_{ab}$
Cost of sampling unit*	c_A	c_B			

*Applies to case of drawing random samples from both frames.

Note that n'_{ab} and n''_{ab} denote respectively the sub-

samples of n_A and n_B respectively which fall into the overlap domain ab. The corresponding means $\bar{y}'_{ab}$ and $\bar{y}''_{ab}$ can only be computed if $n'_{ab} > 0$ and $n''_{ab} > 0$; the same applies to n_a, $\bar{y}_a$ and n_b, $\bar{y}_b$.

4. Formulas for estimation of population totals and means.

In case 1 the estimation problem is reduced to the standard methodology for stratified sampling, whilst in case 4 it will only be possible to estimate population means and not totals. We confine ourselves here to cases 2 and 3. Two approaches leading to identical formulas are possible, viz. (a) the theory of domain estimation, or (b) a special method of weight variables. We use here (b) and introduce the following attributes to units in the two frames:

$$\text{Frame (A)} \quad u_i = \begin{cases} y_i & \text{if } i^{th} \text{ unit is in domain (a)} \\ py_i & \text{if } i^{th} \text{ unit is in domain (ab)} \end{cases}$$

$$\text{Frame (B)} \quad u_i = \begin{cases} y_i & \text{if } i^{th} \text{ unit is in domain (b)} \\ qy_i & \text{if } i^{th} \text{ unit is in domain (ab)} \end{cases}$$

Here p and q are two fixed numbers (to be optimally determined as shown below) with $p+q=1$. We therefore have converted the two frames into two mutually exclusive strata of sizes N_A and N_B by duplicating the N_{ab} units in domain (ab). In stratum (A) there will be N_{ab} units carrying a characteristic $u_i = py_i$ and in stratum (B) there will be N_{ab} units (the 'aliases' of those in stratum (A)) carrying the characteristic $u_i = qy_i$. Clearly Y, the total of the y_i for the original population of $N=N_a+N_{ab}+N_b$ units, is equal to the total U of the u_i for the new population of $N^*=N_a+2N_{ab}+N_b$ units since

$$(1) \quad Y=Y_a+Y_{ab}+Y_b=Y_a+pY_{ab}+qY_{ab}+Y_b=U$$

The standard methodology applicable to the survey designs in Frames (A) and (B) are therefore applicable to obtain estimates of the two stratum-totals (frame-totals) for the variate u_i, their variances and variance estimates. Adding the two will yield the corresponding formulas for the estimation of U and hence Y. To obtain estimates of the population mean $\bar{Y}=Y/N$ apply these formulas to the count variable $x_i=1$ to estimate its total N in analogy to (1) and use the ratio estimate to estimate $Y/N=\bar{Y}$. In case 2 use the device of post stratifying into post strata (a) and (ab) and (b) and (ab) respectively.

5. Formulas for random sampling in both frames in case 2.

We confine ourselves to the simplest case of random sampling in both frames and ignore finite population corrections. In terms of the notation of Schedule 2, using the u-variates in section 4, the (post stratified) estimator of U=Y is given by

$$(2) \quad \hat{Y} = N_a\bar{y}_a+N_{ab}(p\bar{y}'_{ab}+q\bar{y}''_{ab})+N_b\bar{y}_B$$

where the means $\bar{y}_a\bar{y}'_{ab}$ are replaced by $\bar{y}_A$ if either $n_a=0$ or $n'_{ab}=0$, and where likewise the means $\bar{y}_b$ and $\bar{y}_{ab}$ are replaced by $\bar{y}_B$ if either $n_b=0$ or $n''_{ab}=0$. As is well known, under certain restrictions the post stratified estimator Y has a variance approximately equal to that in proportional allocation stratified sampling so that

$$\text{Var } \hat{Y} \doteq \frac{N_A^2}{n_A}\left\{\sigma_a^2 (1 - \alpha) + p^2\sigma_{ab}^2 \alpha\right\} + \frac{N_B^2}{n_B}\left\{\sigma_b^2 (1 - \beta) + q^2\sigma_{ab}^2 \beta\right\} \quad (3)$$

where

$$(4) \quad \alpha = N_{ab}/N_A, \ \beta = N_{ab}/N_B,$$

finite population corrections have been ignored and σ_a^2 σ_{ab}^2 σ_b^2 are the 'within domain' population variances. Assuming a linear cost function

$$(5) \quad C = c_A n_A + c_B n_B$$

the problem of minimizing (3) as a function of p, n_A and n_B subject to (5) leads to the following specification of the 'optimum' two frame sampling plan:

The optimum value of p is given by a solution of the bi-quadratic

$$(6) \quad c_A p^2/c_B q^2 = \frac{\sigma_a^2(1 - \alpha) + \alpha p^2\sigma_{ab}^2}{\sigma_b^2(1 - \beta) + \beta q^2\sigma_{ab}^2}$$

With the help of p the optimum sampling fractions are given by

$$n_A/N_A = c\left\{(\sigma_a^2(1 - \alpha)+\alpha p^2\sigma_{ab}^2)/c_A\right\}^{\frac{1}{2}}$$
$$n_B/N_B = c\left\{(\sigma_b^2(1 - \beta) + \beta q^2\sigma_{ab}^2)/c_B\right\}^{\frac{1}{2}} \quad (7)$$

with c determined to meet the budget (5). In case of multiple roots of (6) substitution of the alternatives in (3) will select the absolute minimum, except in the rare cases in which the minimum is attained on the boundary of the p, $n_A n_B$ space, an occurrence discussed in the literature on optimum allocation in stratified sampling.

Considerable simplifications arise in the important special case in which the A-frame has 100% coverage so that

$$(8) \quad N_{ab} = N_B, \ \beta = 1, \ \sigma_B^2 = \sigma_{ab}^2$$

If these are substituted in (6) the bi-quadratic reduces to the simple equation

$$(9) \quad p^2 = \phi(1 - \alpha) / (\rho - \alpha)$$

where

$$(10) \quad \phi = \sigma_a^2/\sigma_{ab}^2; \ \rho = c_A/c_B; \ \alpha = N_B/N_A$$

In this special case it is possible to compare the variance $\text{Var}(\hat{Y})$ for the optimum design with that of

$$(11) \quad Y' = N_a\bar{y}_a + N_{ab}\bar{y}'_{ab}$$

which is the (post stratified) estimator computed from a simple random sample of size $n_A^* = C/c_A$ drawn from frame A only and requiring an identical budget C. We find for the reduction in variance

$$(12)\quad \mathrm{Var}(\hat{\bar{Y}})/\mathrm{Var}(\bar{Y}') = (1 - \frac{q\alpha}{p\rho})^2/(1 + \frac{\alpha q(1+p)}{\rho p^2})$$

with p^2 given by (9) and ϕ, ρ, α given by (10). The reduction in variance for constant cost C (which is of course also the reduction in cost for constant variance) is tabulated in Tables 1, 2 and 3 for the parameter combinations

$$\phi^{-1} = \sigma_B^2/\sigma_a^2 = 1,\ 4,\ 16$$

$$(13)\qquad \rho^{-1} = c_B/c_A = .01,\ .05,\ .10\ (.1)\ .5, 1$$

$$\alpha = N_B/N_A = .5\ (.1)\ .9,\ .95,\ 1$$

The cost reduction may be considerable. Thus for a characteristic situation $\phi^{-1}=16$, $\rho^{-1}=0.2$, $\alpha=0.9$ the reduction is 0.248, i.e. the survey only costs $\frac{1}{4}$ of the 'A-frame only' survey of the same precision. It should, however, be pointed out that the cost (5) only represents the cost depending on the size of the sample and that the (omitted) overhead cost may be larger for the two frame survey because of its more sophisticated design. It should be noted that when $\sigma_B^2/\sigma_a^2 > 1$ that there are <u>two</u> causes for the cost reduction with the two frame design: The first (which is particularly operative when $\alpha = N_B/N_A$ is near 1) is the lower cost of sampling units in B, the second cause (which is more operative when α is near $\frac{1}{2}$) arises when σ_B^2/σ_a^2 is large as the two frame design gets closer to optimum allocation of expected pooled sample sizes to the domains a and ab = B. This situation may arise when the 'cheap' frame B contains the 'important' units, i.e. units with larger and hence more variable y-values. Here the two frame design operates as a 'screening device'. Note in particular in Table 1, that for fixed ρ^{-1} the variance ratio rises from ρ^{-1} to 1 as α is reduced from 1 to 0. Although only the range $0.5 \leq \alpha \leq 1$ is shown it is clear that the function is not monotonically increasing which is a reflection of the above two causes for variance reduction.

6. An alternative cost function in case 2.

The assumptions involved in the cost function (5) imply that the full cost of c_A/unit is incurred for all n_A units sampled from frame A, and likewise for frame B. It may be argued that it may be possible to ascertain at a lower cost c_A' whether or not the unit belongs to domain (a) or (ab), and not to complete the questionnaires for the n_{ab}' units falling into domain ab = B. Since y_{ab}' would then not be available it will be necessary to put p = 0 and q = 1 in (2) and (3) and to modify (5) by replacing the actual cost C by its expectation E(C) and c_A by

$$(14)\qquad c_A^* = c_A(1 - \alpha) + c_A'\,\alpha$$

With p = 0 and q = 1 fixed the variance Var $\hat{\bar{Y}}$ given by (3) must now be minimized as a function of n_A and n_B only subject to a given expected cost E(C). The mathematical formulas are now completely analogous to optimization for strata allocation. We find for the optimum design

$$(15)\quad \frac{n_A}{n_B} = \frac{\sigma_a}{\sigma_B}\sqrt{1-\alpha}\sqrt{\frac{e_B}{e_A}}\frac{N_A}{N_B} = \sqrt{\frac{\phi(1-\alpha)}{\rho}}\ \frac{1}{\alpha}$$

and for the variance reduction

$$(16)\quad \mathrm{Var}\ \hat{\bar{Y}}/\mathrm{Var}\ \bar{Y}' = \frac{(1 - \alpha\omega)(1 + \frac{\alpha}{\sqrt{(1-\alpha)\phi\rho}})^2}{(1 + \frac{\alpha}{(1-\alpha)\phi})}$$

where now

$$(17)\qquad \rho = c_A^*/c_B \text{ and } \omega = 1 - (c_A'/c_A)$$

The above reduction may be compared with (12). For ω close to 1 (i.e. c_A^* near to zero) (16) may give a smaller value (larger reduction in variance) than (12). However, in many situations (as for example, with the Census example mentioned in 2.) it is not possible to determine in the field to which domain a sampled unit belongs and with a completion of the questionnaire $c_A' = c_A$, $\omega = 0$, the choice of p = 0 is not optimum and the allocations (7) are preferable.

7. Summary of formulas in case 3.

If domain sizes N_a N_b N_{ab} are not known the ordinary stratified sampling formulas applied to the u-variables in the <u>two</u> strata (frames) of size N_A and N_B must be applied. Thus our estimator of $U = \bar{Y}$ is given by

$$(20)\qquad \dot{\bar{Y}} = \frac{N_A}{n_A}\left\{ y_a + py_{ab}' \right\} + \frac{N_B}{n_B}\left\{ y_b + qy_{ab}'' \right\}$$

and its variance by

$$(21)\quad \mathrm{Var}(\dot{\bar{Y}}) = \frac{N_A^2}{n_A}\left\{(1-\alpha)\sigma_a^2 + \alpha p^2\sigma_{ab}^2 + \alpha(1-\alpha)(\bar{Y}_a - p\bar{Y}_{ab})^2\right\} + \frac{N_B^2}{n_B}\left\{(1-\beta)\sigma_\beta^2 + \beta q^2\sigma_{ab}^2 + \beta(1-\beta)(\bar{Y}_a - q\bar{Y}_{ab})^2\right\}$$

The problem of minimizing $\mathrm{Var}(\dot{\bar{Y}})$ as a function of p, n_A, n_B subject to a given cost (5) can again be solved and in the present case leads to the optimum allocation formulas

$$(22)\quad n_A^2/N_A^2 = c\left\{\sigma_a^2(1-\alpha) + \sigma_{ab}^2 p^2\alpha + \alpha(1-\alpha)(\bar{Y}_a - p\bar{Y}_{ab})^2\right\}/c_A$$

$$n_B^2/N_B^2 = c\left\{\sigma_b^2(1-\beta) + \sigma_{ab}^2 q^2\beta + \beta(1-\beta)(\bar{Y}_b - q\bar{Y}_{ab})^2\right\}/c_B$$

with the constant c to be determined from (5). The allocation formulas (22) involve the value of the optimum p (and q = 1 - p) which must again be

determined as the root of a bi-quadratic similar to (6) but not given here. A comparison with an A-only sample design is not appropriate here since the assumption of a frame A with 100% coverage automatically leads to a complete knowledge of the domain sizes from $N_{ab} = N_B$, $N_b = 0$, $N_a = N_A - N_B$ so that the post stratified estimator $\bar{y}$ of section 2 should be used.

8. Planned applications.

The small survey on 'Effect of Industrialization on Farming' mentioned in 2 was not designed in accordance with the optimum formulas of section 5 above, in fact it provides an example of the cost situation discussed in 6. Whilst there was not too much difference in the cost values c_A and c_B there was a considerable difference in σ_B^2 and σ_a^2 for most characteristics, in fact it was decided to sample frame B 100 per cent. It is hoped to incorporate illustrative data from this survey and others using the designs here given at the time of publication.

Table 1. Variance reduction in two frame sampling when $\sigma_B^2/\sigma_a^2 = 16$

Sampling cost ratio c_B/c_A	N_B/N = proportion of population in cheap frame						
	.5	.6	.7	.8	.9	.95	1
.01	.096	.076	.059	.045	.031	.024	.010
.05	.154	.134	.118	.102	.086	.075	.050
.10	.206	.188	.174	.160	.143	.131	.100
.20	.288	.278	.269	.261	.248	.237	.200
.30	.359	.356	.355	.353	.347	.338	.300
.40	.423	.428	.435	.440	.441	.436	.400
.50	.483	.496	.510	.524	.533	.532	.500
1.00	.735	.784	.836	.889	.944	.972	1

Table 2. Variance reduction in two frame sampling when $\sigma_B^2/\sigma_a^2 = 4$

Sampling cost ratio c_B/c_A	N_B/N = proportion of population in cheap frame						
	.5	.6	.7	.8	.9	.95	1
.01	.259	.201	.152	.108	.066	.044	.010
.05	.340	.284	.234	.186	.137	.107	.050
.10	.404	.352	.304	.257	.205	.172	.100
.20	.500	.456	.415	.372	.322	.287	.200
.30	.576	.540	.507	.472	.426	.393	.300
.40	.640	.613	.588	.561	.523	.493	.400
.50	.696	.678	.661	.642	.614	.589	.500
1.00	.900	.914	.932	.953	.976	.988	1

Table 3. Variance reduction in two frame sampling when $\sigma_B^2/\sigma_a^2 = 1$

Sampling cost ratio c_B/c_A	N_B/N = proportion of population in cheap frame						
	.5	.6	.7	.8	.9	.95	1
.01	.571	.477	.379	.276	.164	.101	.010
.05	.656	.573	.482	.381	.260	.186	.050
.10	.718	.645	.562	.465	.344	.263	.100
.20	.800	.742	.674	.589	.475	.392	.200
.30	.857	.812	.757	.686	.582	.503	.300
.40	.900	.866	.824	.765	.676	.604	.400
.50	.933	.909	.877	.832	.759	.695	.500
1.00	1	1	1	1	1	1	1

9. The Bureau of the Census Survey of Retail Stores.

It has been pointed out to us that mention should be made of the 'Sample Survey of Retail Stores' by the Bureau of the Census (1949) which is perhaps one of the largest and earliest instances of the combined use of a list-frame and an area frame. For a description of this survey see Hansen, Hurwitz and Madow (1953) 'Sample Survey Methods and Theory', Vol. 1 (pp. 515-558). This survey carefully avoids the sampling of list-units encountered in the sampled area segments and therefore follows essentially the method described in 6, although the area sample design is multi-stage. In the description of this survey it is not discussed whether the cost c_a' of screening out list units from the area segment warrants their omission and the use of weight coefficients p=0 and q=1 in place of the optimum p and q of Section 5. In fact the method of optimum weight coefficients p and q has to the best of our knowledge never been used. Cost c_a' consists of recognizing a place of business located within segment boundaries making sure that it is identical with an establishment whose address is mentioned on the list and then discarding it from further interview.

An example of a similar type is discussed on pp. 327-8 of the above book, where a table of variance reduction is given for a special situation which corresponds closely to our case 1 in Schedule 2, which follows standard stratification.

SAMPLING FOR WEEKLY TIME SERIES

Max A. Bershad and Harold Nisselson
Bureau of the Census
Washington, D.C.

Introduction

The problem we wish to discuss in this paper is the design of a general purpose repetitive sample survey to provide estimates of a number of statistics. The estimates from the survey on successive occasions are to be used to draw inferences about the population and underlying changes in the population. For this purpose they will be subject to a wide variety of analyses and comparisons, ranging from more or less formal time series analysis to comparisons with preceding weeks, preceding months, the preceding year, and patterns from earlier years. From time to time the data from a number of occasions may be pooled to provide estimates of aggregates or more detailed analyses of the characteristics and changes in characteristics of the population.

There are two interrelated aspects to the problem of survey design in these circumstances. One of these, to which we will give major attention, is the design of a sampling and estimation procedure for each occasion. The various demands on the survey lead to conflicting objectives from a sampling point of view, and the design to be chosen will generally represent a compromise. The other aspect of the problem of survey design is the frequency with which data are to be collected and estimates produced. Given fixed total resources, the more frequent the occasions at which data are to be collected the smaller the sample it is feasible to cover at each occasion. Thus, the question arises as to whether smaller samples should be taken more frequently or larger samples less frequently. This question needs more exploration by survey analysts.[1]

We will be especially interested in exploring the circumstances under which it may be advantageous to use a system of weekly samples even though the expressed primary interest of the survey lies in estimating monthly statistics. The techniques we will discuss are the use of rotating samples and composite estimates [1] [2] [3] [4]. The flexibility of rotating samples provides the survey designer with the opportunity to choose from a number of different plans to meet special circumstances and important demands for data. Among these choices is the opportunity to sample the occasions themselves. We will compare designs based on weekly estimates every week with a design based on a systematic sample of weeks. In the latter case there is a component of variation between weeks that is generally neglected but should be taken into account.

In comparing alternative sample designs we will take as our criterion the sampling variance of estimates from the survey. These estimates will be used in a wide variety of analyses, and variance appears to be one of the best general criteria for all uses.

Rotating Samples and Composite Estimation

Given a repetitive survey, at one extreme the sample at each occasion might be identical. This would be the case of a fixed panel. At the other extreme, the sample at each occasion might be completely independent. This would be the case of complete rotation of the sample. Cases in between represent partial or incomplete rotation of the sample. The pattern of occasions and number of occasions for which any given sampling unit provides data in the survey is the rotation plan. The significance of the rotation plan is that it determines the carryover of information between occasions by identical units.

It is helpful to distinguish the concept of overlap in information between occasions from sample overlap, that is, the extent to which identical sampling units are surveyed at different occasions. If information is obtained from sample units for only one occasion in the interview or report, the overlap in information between occasions is the same as the sample overlap. Overlap in information between different occasions can, however, be achieved without overlap in the sample interviewed if information for more than one occasion can be obtained from sample units in a single interview or report.

To illustrate these concepts, suppose a continuing interview survey based on equal-sized weekly samples of segments with the following rotation plan: Any one segment is in sample only one week of the year and then returns to the sample again in the corresponding week of the following year. At interview, information is obtained from each unit in the segment for each of the two weeks preceding the week of interview. Then the overlap in information between successive weeks is 50 percent and the overlap between corresponding weeks a year apart is 100 percent. This plan is discussed further below.

If we have overlap of information between occasions we are in a position to use data from the past to try to improve current estimates. Suppose we have a continuing weekly survey with overlap in information between successive weeks and we are interested in estimating the level of an item for the current week. We can construct an estimate of level in at least two different ways:

1. By taking the (composite) estimate of level for the preceding week and adding an estimate of change derived from the sampling units for which there is information for both weeks.
2. By making an estimate of level for the given week directly, using the data from all sampling units for which there is information for the current week.

The composite estimate of level for any given week is a weighted average of these two types of estimates. The composite estimate of change between any two weeks is the difference between the composite estimates of level for the two weeks.

The determination of the appropriate weights to use in the composite estimate for a single item is a straight forward problem in optimization. The appropriate weighting to minimize the sampling variance of the composite estimate will however vary from item to item, and between statistics of level and change for a given item, depending upon the pattern of correlations. The determination of an efficient rotation plan is also a problem in optimization. It is not, however, straight forward but has thus far been dealt with only by comparison of specific alternatives. If information for a sampling unit is positively correlated between different time periods, as one might expect a high degree of overlap is desirable for estimating change between time periods. This is not so for estimates of level, or for estimates of aggregates based on the sum of successive sample estimates or analyses for which successive samples are to be pooled. If overlap in information can be achieved in the interview by collecting data for more than one time period rather than by sample overlap, the objectives of estimating both level and change well can be more successfully reconciled than if we must depend on overlapping the sample.

Reasons for Using Rotating Samples with Composite Estimates

It is worthwhile to review in some detail the reasons for using rotating samples with composite estimates compared with the alternatives of a fixed panel or completely independent samples. These will be considered in connection with sampling variances, and measurement and control of response error.

With regard to sampling variances, it will generally be found that where partial rotation is of advantage over a fixed panel, say, a design based on completely independent samples has the same advantage to an even greater extent -- and a similar statement could be made in the opposite direction. Thus, we may think of the use of rotating samples with composite estimation as having some advantages and some disadvantages compared with alternative designs. It combines the possibilities available under either of the two extremes and, although it does not exploit any of them to the hilt, it is fortunate that frequently most of the benefit of a particular feature will be gained even without complete exploitation.

There are certain unique advantages in the measurement of response error that are inherent in the use of rotating samples compared with either a complete overlap of information or no overlap of information.

Sampling Variances

Compared with a fixed panel design rotating samples provide improved estimates of current level. This is so because the composite estimate makes it possible to take advantage of the information in past samples. This raises a further interesting possibility. Ordinarily, a time series is produced point by point as each point in time is reached. Suppose we are willing to revise the current estimates at a later date, say annually, with the object of developing the "best" historical series. The composite estimate can be extended so that the "current" estimate at each point in time takes advantage of the information in samples future to that time as well as in the past samples.

There is a further advantage with rotating samples in the ability to treat more satisfactorily unexpected large units that occur in the sample. When unexpectedly large observations occur in a sample survey a choice must ordinarily be made between accepting the considerable increase in variance they create or of reducing their weight and accepting the resulting bias. The usual advice is to choose the alternative expected to lead to a smaller mean square error. With rotating samples it is possible to improve on this procedure by identifying all large observations in the entire annual sample and including them in the survey for the current time period. The effect of this is to sample large observations at each occasion at a rate k times that of other observations, where k is the ratio of the number of different segments annually to the number in the current survey, and hence to divide their weight in the current estimate by k. While the mechanics of putting this principle into effect may sometimes require considerable ingenuity, the resulting gains can be substantial.

Measurement and Control of Response Errors

We will consider the problems of response error in connection with nonresponse, quality of data for interview cases and measurement of response differences.

Compared with a fixed panel design, the use of rotating samples reduces the burden of reporting on the individual respondent. This can be important in maintaining a high rate of response. Where there is nonresponse, rotating and completely dependent samples both have the advantage over completely independent samples in that an earlier report may be available to permit better imputation or adjustment for the nonresponse cases in the current estimate. This can be of special help for statistics of change between different points in time, since the impact of nonresponse on statistics of change may be more nearly measured by the sum of the nonresponse rates at each of the two occasions than by the individual rates.

Where a response is obtained, previous information may be useful for improving the

current information. This may be realized in any of several ways -- for example, by the use of shuttle forms where the respondent has the opportunity to see his earlier information, or by the application of editing rules leading to follow-up to correct or clarify the information originally reported.

Interview on successive occasions may make it possible to obtain better current information through improved techniques. For example, in pilot studies of the reporting of homeowners' expenditures for alterations and repairs it was found that more precise and complete reporting of smaller expenditures could be obtained by furnishing the interviewer with a copy of the previous response to read to the respondent and asking about expenditures since the previous interview.[6] This has been called a "bounded interview" since the earlier interview bounds the current one.

The opportunity to compare responses for identical units at different points in time with responses for new units that is provided by rotating samples may help to uncover defects in the survey procedure. This was the case in the Census Bureau's Monthly Retail Trade Survey. Data are obtained at interview in that survey for each of the two calendar months preceding the month of interview. It was found that the survey procedure tended to miss stores that had gone out of business during the month preceding interview. This was remedied in part by providing for special field instructions in case of vacant stores.

Where overlap of data is created by asking for information covering more than one time period, the comparison of data for a fixed period obtained from successive panels may point to problems of recall and suggest revisions in the survey procedure. In several household health surveys a recall period of four weeks was used for obtaining reports, with a system of independent weekly samples. Comparison of the data for each calendar week when it was the week preceding the week of interview, two weeks preceding the week of interview, etc., indicated a sharp decline in the level of illness reported with increasing length of recall. In the pilot study for the National Health Survey a two-week recall period was tested by this technique and found to be acceptable.[7]

There is a special aspect to surveying the same unit on more than one occasion, and that is the training or conditioning effect on the respondent. In the case of well-defined item, the effect of repeated interview and questioning of the respondent may reasonably be expected to lead to improved data. When the definition of the item has subjective elements, however, it may be questionable whether the data from later reports are better or worse. Examples can be cited on both sides of the argument. There is, however, ample experience to show that they will frequently be different. With rotating samples there is an opportunity to compare responses of new units in the survey with those of continuing units so that such differences can be measured and explored.

Some Choices among Alternative Sample Designs

We now turn to the specific question of the use of weekly samples posed at the beginning of this paper, and will examine some alternative survey designs (including different rotation plans) from a variance point of view. We will consider various statistics under a composite estimation procedure for two types of items.

We assume the following special conditions for the survey --

(1st) That respondents have satisfactory recall for the last two time periods (weeks) but cannot furnish satisfactory data for a longer period.

(2nd) That primary interest of the survey lies in estimating either the total value of an item for four time periods (month) or the average value of the item for four time periods and in the changes over time in these totals and averages.

(3rd) That the estimate of total or average for the item is to be published at the end of every four time periods on a timely basis. Changes in level are to be derived from the published totals or averages.

Many rotation patterns are available for use under these circumstances. Four possible plans of rotation will be described here. In three plans a predesignated number of interviews are made every week. These plans have a weekly overlap of 50 percent in information, since they involve asking the respondent for data in the last two weeks, but they differ from one another in either their monthly overlap or yearly overlap. In the fourth plan, a systematic sample of weeks is taken, one week from each month; four times the predesignated number of interviews are concentrated in the sample week and the respondent is asked for data for only the last week.

Description of the Rotation Plans

The first plan considered is the "50-75-50 Plan." This plan is characterized by a 50 percent overlap in information from week to week; a 75 percent overlap in sample from month to month, and a 50 percent overlap in sample from year to year. A respondent on his first interview furnishes separate data for the last two weeks; he is interviewed three more times, the interviews being spaced by four-week intervals, before he has an eight month respite. After this he is interviewed four more times at monthly intervals, so that the total number of interviews with this respondent is eight.

The interview pattern is sketched in Diagram I, where each line represents a different week and each column represents a different sample person (or group of persons). Different sample persons are designated in the diagram by different letters. As an illustration, the sample person or sample group designated by the small letter "m" is interviewed in weeks t, t-4, t-8, t-12, t-52, t-56, t-60, and t-64.

The second plan is the "50-0-100 Plan," indicating a 50 percent weekly overlap, 0 percent monthly overlap, and 100 percent yearly overlap. Each sample element is interviewed just one a year, in every year.

The third plan is the "50-50-100 Plan," and indicates 50 percent weekly overlap, 50 percent monthly overlap, and 100 percent yearly overlap. After the first interview, a sample respondent is interviewed four weeks later, is out of the sample for 10 months, then returns for two interviews, is out again for 10, etc. Each sample element is interviewed twice a year.

The fourth plan is the "X-75-50 Plan," under which interviews are conducted only in a systematic sample of weeks, one from each month. On interview, the sample person reports for the last week only and not for the last two weeks. The sample person is interviewed in the sample week in four consecutive months; he is out for eight months and then in again for four months. Thus, eight interviews are conducted for each sample element. As mentioned previously, the number of interviews in the sample week under this plan is equal to the number of interviews conducted in a month for one of the previous three rotation plans.

Results

The tables below compare the variances of estimates of different statistics for two different items as derived from estimating the weekly level of an item by means of the "composite" estimator employing a weight of 0.5. In the first three plans advantage is taken of those responses for the week being estimated obtained from the sample elements during interviews made in the subsequent week. Except for a factor of 1/n for sample size, the number 4.00 represents the variance of a simple unbiased estimate of weekly level of an item from an independent weekly sample of size n; or equivalently the number 1.00 represents the variance of the average of four such weekly levels during a month.

The following conclusions can be drawn from the tables:

(1) In general, the different plans have different abilities to estimate the different statistics. For level those plans are best which utilize more different and distinct respondents. In alternate terms those plans are best for level which involve fewest repetitions of interviews with the sampled elements.

(2) In general, for change between two time periods, those plans are best which have the highest overlap between the given time periods. This is somewhat conditioned by the fact that the plan must not lose too much on level because of repetitive interviews.

(3) That there may be purposes for which individual weekly levels themselves and changes in these levels have satisfactory variances even if the objective at the outset might have been to provide monthly averages or totals of the weekly levels. (This conclusion follows from the fact that the weekly levels generally have less than twice the variance of the monthly average.)

(4) Dependent on the size of the between week variance contribution, (and the size of sample employed) the "X-75-50 Plan" may result in poorer estimates of monthly averages or totals than any of the other three plans which do not involve sampling of weeks.

Footnote

[1] For a discussion of this question in the context of statistical quality control see [5].

References

[1] Hansen, M. H., Hurwitz, W. N. and Madow, W. G., "Sample Survey Methods and Theory," Vol. II, Ch. 11, Sec. 8-9, John Wiley and Sons, Inc., 1953.

[2] Eckler, A. R., "Rotation Sampling," Annals Math. Stat., 26 (1955), 664-685.

[3] Woodruff, R. S., "The Use of Rotating Samples in the Census Bureau's Monthly Surveys," Proc. Soc. Stat. Sec., Amer. Stat. Assn., 1959, 130-138.

[4] Hansen, M. H., Hurwitz, W. N., Nisselson, H. and Steinberg, J. "The redesign of the Census Current Population Survey," Journ. Amer. Stat. Assn., 50 (1955), 701-719.

[5] Weiler, H., "The Use of Runs to Control the Mean in Quality Control," Journ. Amer. Stat. Assn., 48 (1953), 439-446 (and Errata, 51 (1956) 652).

[6] Neter, J. and Waksberg, J., "Measurement of Nonsampling Errors in a Survey of Homeowner's Expenditures for Alterations and Repairs," Proc. Soc. Stat. Sec., Amer. Stat. Assn., 1961, 201-210.

[7] Nisselson, H. and Woolsey, T. D., "Some Problems of the Household Interview Design for the National Health Survey," Journ. Amer. Stat. Assn., 54 (1959), 69-87.

Table I

COMPARATIVE VARIANCES WITH FOUR DIFFERENT ROTATION PLANS FOR ESTIMATES OF LEVEL AND CHANGES OVER VARIOUS TIME PERIODS: A HIGH CORRELATION[a] ITEM

Statistic	Estimate of statistic based on -						
	Average of 4 weekly estimates			Individual weekly estimate(s)			
	50-75-50 Plan	50-0-100 Plan	50-50-100 Plan	50-75-50 Plan	50-0-100 Plan	50-50-100 Plan	X-75-50 Plan
Level.....	0.92	0.70	0.85	1.27	1.14	1.23	0.83[b]
Difference between two levels:							
1 month apart.....	0.59	1.05	0.82	1.08	2.09	1.46	0.55
1 year apart.....	1.23	0.49	0.57	1.75	0.88	0.94	1.14
1 week apart.....	c/	c/	c/	0.66	0.73	0.69	d/
3 months apart....	1.51	1.40	1.68	e/	e/	e/	1.22
6 months apart....	1.85	1.40	1.69	e/	e/	e/	1.58

a/ A high correlation item is one for which $\rho_1 = .95$, $\rho_3 = .85$, $\rho_4 = .80$, $\rho_5 = .75$, $\rho_7 = \rho_8 = \rho_9 = .70$, $\rho_{11} = \rho_{12} = \rho_{13} = .65$, $\rho_{51} = \rho_{52} = \rho_{53} = .70$, $\rho_{47} = \rho_{48} = \rho_{49} = \rho_{55} = \rho_{56} = \rho_{57} = .65$, $\rho_{43} = \rho_{44} = \rho_{45} = \rho_{59} = \rho_{60} = \rho_{61} = .60$, $\rho_{39} = \rho_{40} = \rho_{41} = \rho_{63} = \rho_{64} = \rho_{65} = .60$; where ρ_t is the correlation over a t week period.

b/ The estimate of weekly level can be made for but one week during the month. The estimate for the week can be used as the estimate for the average of the 4 weeks from which the week was sampled, in which case the variances in this column must be increased to reflect the between-week variance.

c/ Inapplicable

d/ Not possible

e/ Not available

Table II

COMPARATIVE VARIANCES WITH FOUR DIFFERENT ROTATION PLANS FOR ESTIMATES OF LEVEL AND CHANGES OVER VARIOUS TIME PERIODS: A LOW CORRELATION[a] ITEM

Statistic	Estimate of statistic based on -						
	Average of 4 weekly estimates			Individual weekly estimate(s)			
	50-75-50 Plan	50-0-100 Plan	50-50-100 Plan	50-75-50 Plan	50-0-100 Plan	50-50-100 Plan	X-75-50 Plan
Level.....	1.14	0.79	1.09	1.98	1.69	1.95	1.00[b]
Difference between two levels:							
1 month apart.....	1.28	1.27	1.41	2.93	3.21	3.16	1.16
1 year apart.....	1.83	0.91	1.37	3.40	2.23	2.82	1.62
1 week apart.....	[c]	[c]	[c]	1.81	2.02	1.82	[d]
3 months apart.....	2.10	1.57	2.16	[e]	[e]	[e]	1.73
6 months apart.....	2.27	1.57	2.17	[e]	[e]	[e]	1.96

a/ A low correlation item is one for which $\rho_1 = .70$, $\rho_3 = .56$, $\rho_4 = .50$, $\rho_5 = .44$, $\rho_7 = \rho_8 = \rho_9 = .40$, $\rho_{11} = \rho_{12} = \rho_{13} = .30$, $\rho_{51} = \rho_{52} = \rho_{53} = .50$, $\rho_{47} = \rho_{48} = \rho_{49} = \rho_{55} = \rho_{56} = \rho_{57} = .45$, $\rho_{43} = \rho_{44} = \rho_{45} = \rho_{59} = \rho_{60} = \rho_{61} = .40$, $\rho_{39} = \rho_{45} = \rho_{41} = \rho_{63} = \rho_{64} = \rho_{65} = .30$; where ρ_t is the correlation over a t week period.

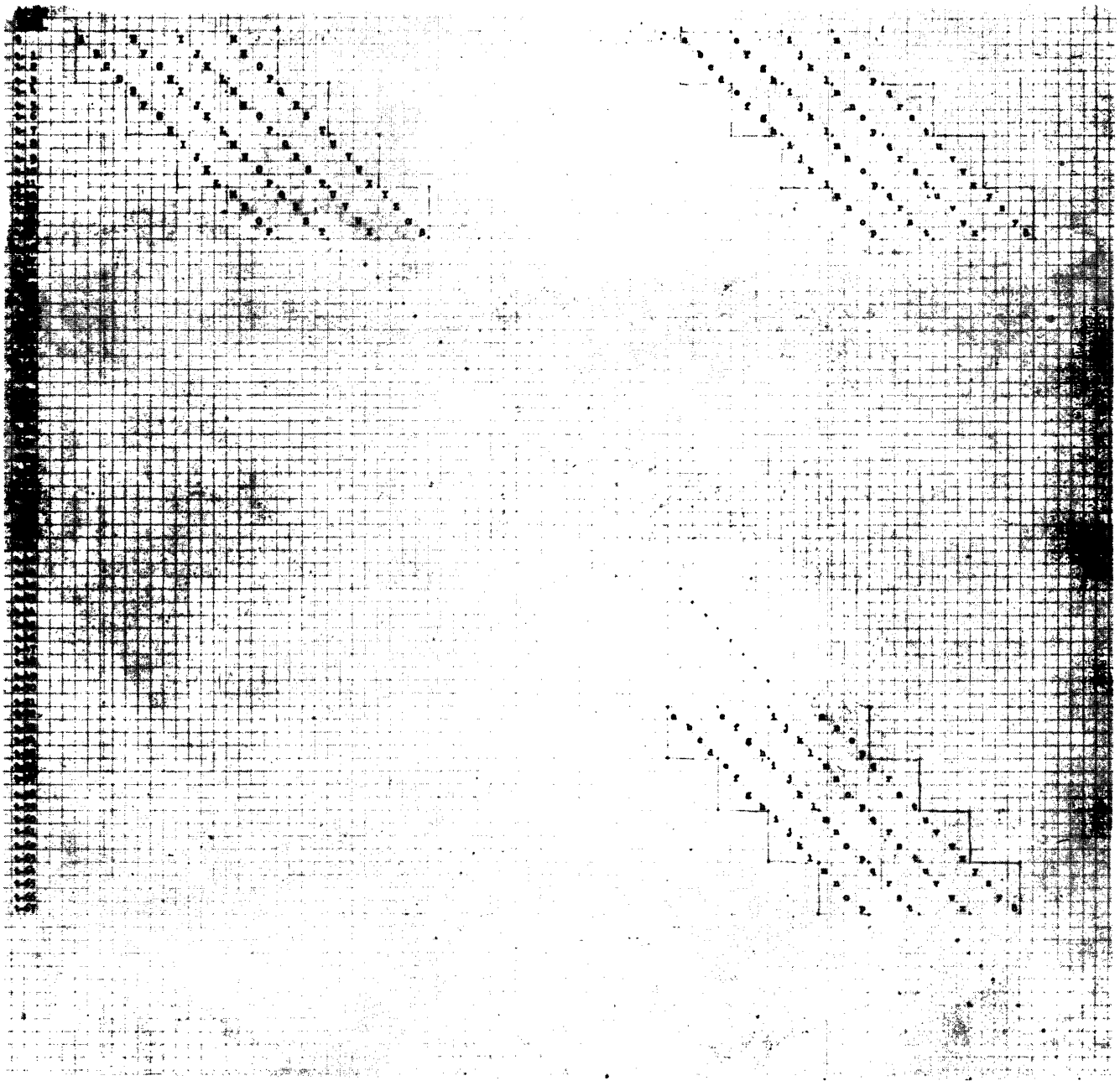
b/ The estimate of weekly level can be made for but one week during the month. The estimate for the week can be used as the estimate for the average of 4 weeks from which the week was sampled, in which case the variances in this column must be increased to reflect the between-week variance.

c/ Inapplicable

d/ Not possible

e/ Not available

Diagram I -- R O T A T I O N P A T T E R N F O R T H E "50-75-50 PLAN"



IX

METROPOLITAN GROWTH

Chairman, Roye L. Lowry, Federal Statistics Users' Conference

Page

Urban-Regional Economic Analysis: Concepts, Measurement and Data - Wilbur R. Thompson, Wayne State University..... 216

Panel Discussion:

Benjamin Chinitz, University of Pittsburgh..... 225

Charles Tiebout, University of Washington..... 227

Kirk R. Petshek, University of Wisconsin..... 228

URBAN-REGIONAL ECONOMIC ANALYSIS: CONCEPTS, MEASUREMENT AND DATA

Wilbur R. Thompson, Wayne State University

This paper is a condensation of the first half of a manuscript written over the past academic year while enjoying a research fellowship sponsored by the Committee on Urban Economics of Resources for the Future. The parent piece is intended as a stimulus to research and a guide to research priorities in the young and rapidly growing field of urban economics. Because even the unabridged version is a potpourri of concepts, hypotheses, methodology and data references, leaving the task of separating the wheat from the chaff to reader-critics, a condensation runs the risk of becoming an almost incomprehensible kaleidoscope of (half-digested) ideas. The saving grace lies in the fact that the discussants have had access to the unabridged version and, more important, they are three of the most knowledgeable students of the subject matter in the country.

The material summarized here might be characterized as urban-regional economics -- the economics of the small-area, open economy, highly specialized and subject to easy entry and exit of labor and capital. The questions are the conventional ones: the factors and processes that bear of the level, distribution, stability and growth of income and employment. The conceptual framework is also quite orthodox: export and investment multipliers, income and price elasticity of demand, market structure as a determinant of firm behavior. The principal distinguishing feature is the heavy emphasis on location and spatial pattern -- the city in the national system of cities. (The second half of the parent work, not included, begins instead with the "awful mess" of the city planner and administrator and is more concerned with intra-urban phenomena -- traffic, taxes, slums, and the like.) Neither the omission of footnotes nor the personalization of the content should be interpreted as a broad and sweeping claim to originality, although the author hopes he has rendered a useful service by marshalling a modestly original collection of concepts in such a way as to suggest promising and even exciting lines of enquiry to others, that is, carried forward the business of the Committee on Urban Economics.

Level of Income

Most of the legacy with which the urban economist begins his comparative analysis of urban income levels is the only tenuously relevant interstate income analyses of Hanna and associates and Perloff and associates. These interstate studies of per capita income hold three principal lessons for us: first, interregional (interstate) income differentials are largely wage differentials; second, demographic patterns (age, sex) vary significantly between states leading to significantly different labor force participation ratios that appear to account for as much as 20 to 30 per cent of the observed differences; and third, the major proportion of the interstate income differentials is attributable to rural poverty coupled with widely varying urban-rural mixes between states -- both directly, and indirectly through demographic variation -- that the extension of the interstate findings to inter-urban comparisons is highly tenuous. (An interstate per capita income range of roughly 3 to 1 is compressed to about 1.75 to 1 between standard metropolitan areas.) Two conclusions are inescapable: (1) urban-regional income analysis must not be inferred from superior state data but must be conducted at the metropolitan area (labor market) level, albeit with poorer data and (2) nation-wide inter-regional income inequality is less critical to urban economics than to rural-farm economics (intra-area spatial patterns of income distribution are, however, especially critical in the urban context, as we shall see).

Local Labor Markets: Mobility and Wage Determination

It is, in fact, as a labor market -- an income generating entity -- that the economist first establishes rapport with the urban area as a significant unit of economic analysis -- comparable to the time-honored industry. The first steps in local labor market analysis can be made on the proven ground of industry-mix standardization. The conventional technique is to derive a hypothetical wage structure, imputing national wage rates to the locality, industry by industry. An expected local average wage is derived to which the actual (observed) local wage can be compared, leading to inferences about relative local productivity and "earnings rates" (local rate for a given job). But all this is complicated (and enriched) by the need to complement industry and occupational-mix standardization with an area effect (or wage "rollout").

Specifically, if the dominant local industry (steel in Pittsburgh, automobiles in Detroit) is a high wage industry, this probably raises wages generally throughout the local economy. One would guess that shoe clerks (barbers) make more in steel, auto and chemical towns than in textile and shoe towns, especially if shoe clerks (barbers) are more mobile between occupations and industries within the area (more likely to take jobs in nearby factories) than clerks all over are mobile between towns (likely to migrate to retain their trade). To ignore the area effect is to run the risk of inferring that the average earnings rate in areas with high wage export industries is above the national average (due, say, to superior local productivity), when it really is higher simply because local service wage rates have been imputed at the national rate (related to an imaginary national labor market) rather than at a distinctively higher local rate reflecting the wage rollout effect. One of the most valuable statistical projects would be to build an

econometric model of the wage determination process in local labor markets with especial reference to labor mobility as between industries, occupations and areas.

Productivity, Price Power, Politics and City Size

But a labor mobility study is largely a study of the supply side of the labor market, the demand side remains to be considered. The demand for labor is a function of its productivity, interpreted broadly to include both the physical output per worker and the price at which that output is sold. The value productivity of local labor can be traced back to the amount of capital per worker, market structure, political factors, and city size, among other factors.

While data on capital stock per worker is not available on any systematic or comprehensive basis, least so for urban areas, the Censuses and Annual Surveys of Manufactures do report annual expenditures on new plant and equipment for standard metropolitan areas at the "three digit" SIC level in Census years and at the "two digit" level in the interim Survey years. By interpolating between Census year benchmark figures, plant and equipment expenditures could be cumulated for SMAs at the three digit level for the full period 1947 to date. If some reasonable assumptions could be made about the size of the beginning stock of capital and the appropriate rate of depreciation (e.g., beginning stock proportionate to additions, or nearly valueless, or inversely related to the trend -- extrapolating backward in time), we might put together some rough but usable extimates of capital-to-labor ratios for the Census metropolitan areas. (And metropolitan areas come close to defining local labor markets.) The human side of labor productivity might be explored, simultaneously, with Census and other data on educational background (as a proxy for skill) and age and other demographic data (as proxies for energy and effort).

But the relationship between the price power of the leading export industries of an urban area and the level of local income is probably at least as critical as capital, skill and physical productivity. High (administered) prices enrich an area only to the extent that they are translated into higher wage rates (the wage component of value added becomes practically the whole amount in the case of the absentee-owned corporation -- the branch plant of the national corporation). Thus a secure and aggressive local union provides the sufficient condition for local affluence -- at national expense.

Certainly, firm price power and union aggressiveness will be difficult to quantify, but even simplistic surrogates such as degree of concentration in the industry (perhaps coupled with the trend of demand for the product) and per cent of total industry employment unionized could take us a good part of the way. And this empirical hurdle must be cleared for our tentative conclusions about physical productivity and local income level will be most tenuous in the absence of some real feeling for the role of local export price power. Who can say that a high capital-to-labor ratio has produced high local incomes through physical productivity when an equally plausible explanation is that the large investments per worker imply difficult entry into the industries of local specialization, with the fruits of local monopoly power partly shared with local labor.

The value productivity of the local labor force may also be enhanced by political intervention at the Federal level. The support of farm prices clearly enriches rural service centers and also the many manufacturing towns which supply farm equipment, fertilizer and the like (e.g., Waterloo, Iowa). Or the Federal intervention may be on the cost side, as through favorable depletion allowances; Tulsa and Houston have median family incomes that run about 20 per cent above average for similar size cities of their region. (These latter two towns combine the union-oligopoly effect sketched above and the political productivity effect at issue here.) Probably the most pervasive Federal political impact on local living standards is through tariffs, evidenced by the powerful and widespread resistence to our periodic efforts to reduce them.

Attempts have been made to "explain" urban income differentials by correlating them with city size, the inference being that bigger cities create a distinctive working environment, or perhaps they assemble a population group of higher economic productivity -- younger adults, the highly educated, the ambitious. The Hanna-Mansfield studies suggest a modest tendency for income level to rise by about 25 per cent from the lowest of eight city size classes to the highest. Some of the supporting work further suggests that variation in demographic-mix with city size is one significant factor -- the larger cities have a greater proportion of middle-aged persons in the prime of their earning power. Renewed effort on city size and income level with the 1960 Census data has a high priority, if for no other reason than the obligation of the economist to contribute to the city planners problem of optimum city size -- holy grail or not. City size decisions should not be made solely on aesthetic grounds.

From Money Income to Real Income

If the export sector is the (urban) family breadwinner, then the local service sector is analagous to the housewife who bears primary responsibility for the efficiency of the (urban) household as a consuming unit. Specifically, a big pay check won from the local manufacturing plant can be at least partly offset by a poorly managed transit system, an inefficient municipal water plant or a collusive and unenterprising retail trade industry. We know literally nothing about the relative cost of living between cities, a deplorable statistical gap.

The efficiency of the local service sector is probably a function of the spatial position of the urban area relative to natural resources and in the national system of cities, and also the population size of the city region, among other factors. An isolated city is forced to become

self-sufficient at inefficiently low levels of output in many services, or do without; a comparison of living costs between, say, Spokane and South Bend, may shed some light on the cost of enforced self-sufficiency. A high degree of local self-sufficiency may, of course, also be achieved more gracefully and economically through growth, as the larger local market pushes more and more activities across the threshhold level of economical local production. This lowers the cost of previously imported goods and/or increases consumer satisfaction by broadening the range of locally available goods. The quantification of the value of greater selection will obviously be difficult.

City size has an ambivalent effect on urban efficiency. Increasing the number of competitors (e.g., department stores) tends to lower cost and stimulate innovation in the private sector, but population expansion tends to spill across the central city limits, creating the political fragmentation of a natural economic entity and raising the cost or lowering the quality of public services by destroying efficient monopoly. Paradoxically, then, city size breaks local monopolies in both the private and public sectors, effecting gains from competition in the former and the loss of internal and external economies of scale in the latter.

The translation of inter-urban money income differentials into real income differentials calls for an extension of economic base theory. An urban area not only earns its money wage rate but must buy it back as the cost of goods and services produced by local service industries reflects the local wage structure. If, for example, automobile and steel workers have higher than average wage rates, and if wage rollout is substantial, so that barbers and bus drivers in auto and steel towns earn more than their counterparts elsewhere, then the cost of living will be higher in these high wage towns, unless labor productivity of these latter workers is commensurately higher.

Suppose about one-third of total local economic activity (employment) is engaged in producing for export, then about one-third of total local income is used to purchase value added outside the local economy, that is value which is added at a national average wage level, and two-thirds is spent for value added in local production and reflects the distinctive local wage structure. If the local export (auto) wage runs about thirty per cent above average and raises local service wage rates to, say, twenty per cent, above average, then auto town households will face a price structure averaging about 13 per cent (2/3 x 20 per cent) above the national average, and perhaps 25 per cent more than households in a low wage (textile) town. Thus, the 23 per cent (weighted average of one-third "auto" workers and two-thirds "clerks") higher money income yields only a 9 per cent higher real income (1.23 ÷ 1.13).

We began this section by observing that inter-urban money income differentials were substantially less than interstate differentials (primarily due to the urban-rural mix differences). We close by observing that the wage rollout effect may further narrow the real income gap between rich and poor cities. This suggests that the role of the Federal authority in urban affairs may be less as an agent of regional redistribution of income (a la the farm program), and more as an agent of urban efficiency and interregional redistribution of population and capital in accordance with national objectives.

Distribution of Income

The distribution of income is even more the neglected stepchild of urban economics than of economics in general. With almost literally no precedent to guide or prejudice us, an integrated three-part framework is suggested as a point of departure. Beginning with the urban area as a whole (the local labor market), the basic amount of inter-personal income inequality can be measured and analyzed. Inter-personal income inequality can then be traced one step farther down into the political subdivisions of the metropolitan area. Finally, the variation in per capita income (and wealth) between the various political subdivisions can be computed and examined for its political-economic implications.

Area-Wide, Inter-Personal Inequality

The Decennial Census of Population data on family incomes lends itself rather well to this schema. The minor inconvenience of working with a frequency distribution with an open-end upper income class ($10,000 and over in the 1950 Census, happily raised to $25,000 and over in the 1960 Census) can be circumvented by resorting to one of the simpler measures of relative variation - using quartiles or deciles. Employing the inter-quartile variation, this author analyzed the 1950 Census family income data for 151 standard metropolitan areas and found that 46 per cent of the inter-area variation in income inequality could be "explained" by the proportion of total local employment engaged in manufacturing. The figure rose to 61 per cent when non-white as a proportion of total population was added in a multiple correlation and regression analysis. A high degree of income inequality is apparently characteristic of non-manufacturing urban economies with a high proportion of non-white population. The egalitarian hand of the union in manufacturing areas and racial discrimination in the labor market in direct relationship to the proportion of non-whites is strongly suggested by these very preliminary findings.

The opportunity for extension of this work is almost unlimited. A host of intriguing hypotheses come quickly to mind. For example, one would guess that income inequality should increase with city size because, on the supply side, large cities attract not only the most gifted workers but also enough of the most unskilled and impoverished to hold the lower end of the spectrum in place. On the demand side, moreover, large cities host the executive suite of the large corporation with its palace guard of high-priced lawyers, research directors and financiers, matched with openings for an army of cleaning women, elevator operators and parking lot attendents. Although second jobs and second earners in the household, probably more

characteristic of the big city, tend to equalize family incomes, the possibilities of earning a very large proprietorship income exists only in big cities, probably more than offsetting the former effect. All in all, the big city would seem to produce an enviornment especially conductive to greater income inequality -- a hypothesis offered to some uncommitted doctoral student.

From Residential Patterns to Public Finance and Welfare

The basic degree of income inequality exhibited by the local labor market -- the metropolitan area as a whole -- is an inequality that may be blunted or sharpened, in social implication, depending on the internal spatial pattern it assumes, relative to the spatial pattern of local government. Because houses of similar value (and households of similar income) tend to cluster, small political subdivisions can easily encompass only the rich or poor, whereas large political subdivisions must almost of necessity mix both rich and poor to some extent. A metropolitan area that is highly subdivided (politically fragmented) will tend to exhibit a high degree of inter-governmental variation in per capita income and wealth.

Local public finance problems are made greater under heavy political fragmentation not only by disparities in fiscal capacity, but also because public service needs tend to be greater among the poor. Urban life, moreover, is much more complex and interrelated than rural life (external economies and diseconomies are more significant), and "free" local public services could come to be the principal instrument of income redistribution in the coming decade, as local public budgets grow and the progressivity of the Federal income tax is reduced -- but only in fiscally balanced political entities. Perhaps no urban economic question is more pressing than the quantification and rationalization of intra-metropolitan-inter-governmental income and wealth spatial patterns.

We can broaden our vision of the implications of the spatial distribution of income to local public finance by asking this question: Should we enlarge the political subdivisions of the metropolitan area or should we push income-mixing down to finer levels of areal subdivision, or both, to equalize local fiscal capacities and public service needs? Through zoning and other land use controls, public investment and other powers, we might exercise an appreciable, even if not overwhelming, influence on land use patterns. Substantial income mixing must be achieved at least as low as the level of major political subdivision of the metropolitan area if local government is to remain an autonomous, effective and viable institution.

While we are a long way away from achieving mixed income neighborhoods, at what intermediate level might we carve out the balanced income "community" that is probably the key to sophistication in our public welfare policy and to economic democracy in our schools. Wanting to help the less fortunate is not enough, we must know how, and the "culture of poverty" is not learned at great distance -- social or spatial. Further, school segregation follows from residential segregation -- color and income both -- and cross-hauling students is not the answer. How to achieve the big city equivalent of Conant's ideal small town comprehensive high school? The economist has much to contribute to the city planner's and urban administrator's work of rationalizing the spatial pattern of the city, and the Censuses of Population and Housing are exceedingly rich resources whose surfaces have barely been scratched.

Stability of Income and Activity

"Seasonals" and Efficiency

Urban economic instability manifests itself in at least three forms: seasonal patterns, business cycle fluctuations and shifts in growth trends, each of which is distinctive in the urban context. The concept of the "seasonal" is extended here to include all kinds of short period, repetitive oscillations in economic activity: the conventional monthly seasonal in department store sales, the weekly "seasonal" in the use of recreational facilities and the hourly "seasonal" in traffic flow. Seen this broadly, urban seasonals are the crux of public investment and urban efficiency, are a broad rubric covering the pervasive problem of alternating periods of congestion (too little capacity) and slack (idle capacity). In efficiency terms, the urban seasonal problem is to find the point of optimum trade-off between economizing on capital with peak period delays and supporting the extravagance of accomodating peak period traffic with no delays.

Two strategies come quickly to mind. First, we might devote more effort to stabilizing the fluctuations in demand that plague us; we might stagger working hours to eliminate traffic congestion, stagger vacation days and "week-ends" during summertime to stretch our limited outdoor recreation facilities farther, and experiment more courageously with the school year. Certainly we pay to smooth out demand -- staggering working hours relieves traffic congestion at the cost of congestion in communication channels as the period of inter-personal communication and exchange for business purposes is contracted -- but a rational decision can not be made until the net costs and benefits are struck. Second, we might create greater flexibility in the facilities which supply the seasonal demands. Local government employees may be made to double in brass -- combination policemen and firemen -- and public capital can be made to serve more than one use -- the high school auditorium as a community center and the classrooms as the workshops in which vocational re-training will transform square pegs to fit the new round holes of the automated society.

Local Business Cycles: City Size and Industry-Mix

Nowhere is the constraining influence of an economics largely developed in a national context more evident than in the matter of economic

stability. Because business cycle analysis has long been a main stream of thought, much more attention has been paid to the local business cycles than to either local seasonal or growth instabilities, despite the fact that there is probably less the locality can do about its cycle instability than either of the other kinds. But before supporting that assertion, let us examine the nature of the local cycle a bit, for it is a very interesting phenomenon even if it is not especially amenable to local treatment.

The relation between the size of an urban economy and its cyclical stability is of interest if only because a number of deductive traps are hidden here for the unwary. A too literal application of multiplier logic, for example, could lead one to conclude that big cities are more cyclical unstable than small ones. The larger the urban area, the greater the proportion of local income spent for locally produced goods and services, the lower the local marginal propensity to import, the greater the local multiplier. (This follows from the fact that large cities tend to produce a greater range of goods and services.) But to quickly conclude that the higher multiplier of the larger urban area implies greater cyclical instability would be to fly in the face of common sense and casual observation.

Briefly, the explanation of the seeming paradox is simply that the complement of a low marginal propensity to import is a relatively small export sector, therefore the larger multiplier of the big city is multiplied against a relatively small primary oscillation. In its strictest application, economic base theory assumes that the direct relationship between city size and the size of the multiplier is exactly offset by the inverse relationship between city size and the relative size of the export sector, so that the relative change in total activity is always proportionate to the relative change in the export base. The tentative conclusion of the speaker is that elaborate efforts to measure the local multiplier do not rate a high priority.

Not only do we find no deductive presumption favoring a systematic functional relationship between city size and cyclical stability operating through the multiplier effect, but neither is there much of a case for stability through size by way of some functional relationship between city size and industry-mix. Certainly, larger size does bring greater industrial diversification, but cycle peaks and troughs of the various industries tend to coincide in timing (otherwise there could not be an aggregate cycle, of course). Therefore, an urban area cannot match industries with complementary cycles -- as one might combine coal and ice seasonals -- and random industrial diversification, the usual consequence of growth, will lead toward a cycle pattern approximating the national one -- an improvement for Detroit and Pittsburgh but not for Washington, D. C. The hypothesis advanced here, then, is that large urban economies tend to have diversified industrial structures and replicate the national degree of cyclical instability; smaller urban economies exhibit a much greater range of cycle instability; smaller urban economies exhibit a much greater range of cycle instability by virtue of their heavier specialization -- in both the more unstable and the more stable industries.

As we move down from the national economy to the smallest local economy in our income analysis, investment declines and exports rise in relative importance as the prime mover of the business cycle. Can we not deduce in reverse, and in a more dynamic framework, that as the local economy grows in size and becomes more self-sufficient, investment comes to rival exports as an income determinant? But if we open the door to investment as a second autonomous local cyclical force, we might as well open it wide and recognize autonomous shifts in the consumption function. With almost all purchases of producers durable equipment coming from outside, the business cycle in the small urban economy (50,000 population and under) can be characterized as substantially an export cycle with little violence to reality -- some minor adjustment might be made for the local construction cycle.

In the very large urban economy (500,000 or more), exports probably still account for the largest absolute amount of autonomous variation in total local income, but with exports down to perhaps one-third of total local activity and with some local investment purchasing locally produced durable equipment and with a higher local value added component in local construction (more of the building materials produced domestically), local investment expenditures begin to command attention. Not only are local investment expenditures more likely to remain within the local economy with increased city size, but such spending is probably more volatile, on the average, than export earnings which originate in a mixture of producers goods and the generally more stable consumers goods. Thus local investment fluctuations (counted only to the extent that they are reflected in local value added and income generation) almost certainly merit recognition as a secondary force in the local business cycle.

Moreover, with local consumption of local production rising with city size to account for as much as sixty to seventy per cent of total local activity (and income) in these half million population urban areas, even minor autonomous shifts in the "local-consumption-of-local-production function" (e.g. two per cent) could account for as much as 10 to twenty per cent of the total autonomous variation in spending activating the local business cycle. We may, on diligent search, find local economies (a) specialized in cyclically stable exports (e.g., cigarettes, insurance services), (b) expanding facilities rapidly, and (c) sufficiently remote from other areas that they are relatively self-sufficient in consumer goods, such that autonomous shifts in investment and in their local consumption function outweigh export impacts. Are the most unique local cycles, especially in reference to timing, in the more rapidly growing and/or out of the way places? Some unusually attractive opportunities for building econometric models are evident in this virgin territory. As a beginning, an annual manufacturing employment time series for the period 1946 to date can be pieced together from

the Census and Survey of Manufactures, OASI and Labor Market data for about sixty of the larger metropolitan areas -- those with forty thousand or more manufacturing employees.

With the business cycle primarily a demand phenomenon, the income elasticity of demand for the area's principal exports becomes the key concept. The durability (postponability) and use (producers or consumers goods) of local exports are probably the two most critical considerations, with durable producers goods producing the heaviest cycle impact and non-durable consumers goods the least. The quality line is also of interest, with deluxe line products the more income elastic and volatile. With our postwar business cycles increasingly moderated by ever greater built-in stability, the durability of local exports may come to play the dominant role in the new depression-proof, inventory-recession-prone business cycle.

The structure of the product market in which the area exporters sell may also merit attention. What is the net balance of having oligopolistic export industries? Administered (rigid) prices place the full brunt of a decline in demand on production and employment (in contrast to a soft competitive price which sags to encourage consumption). This is offset by the greater depression staying-power of oligopolies, rich in reserves built up via administered prices in the preceding prosperity period, and the easier access to emergency credit of the big, name corporation. In short, oligopolies may cut back output and employment more in early response to a recession decrease in demand, but they are less likely to fail and eliminate all employment. But, if depressions are a relic of the past, the depression staying-power advantage of the oligopoly may be less important than the tendency of rigid prices in recessions to destabilize local employment.

The internal organizational character of local export firms may be crucial in many cases. A local plant which is the branch plant of a multi-plant firm may be operated at full production during a national recession or have its output cut back drastically, depending on whether its the newest (lowest cost) or oldest facility, the plant producing the luxury or economy line of the product, the best or worst located, and other factors. The range (standard deviation) of cyclical possibilities is greatly widened in an urban economy based on branch plants over that which it would experience with single plant firms. The proportion of total automobile production accounted for by the (older plants of the) Detroit area varies directly with total sales and the national business cycle, suggesting that this phenomenon is not confined to small local economies. The financial structure of local firms -- the proportion of debt to equity capital -- and the willingness of local bankers to extend emergency credit are local cyclical factors which are not beyond hope of comparative measurement. All in all, there is a supply side to the local business cycle which cannot be ignored, even though the cycle is primarily a demand phenomenon.

Growth Instability: Migration and Public Investment

Trends are only straight lines because we draw them that way. If we knew the true trend we might measure growth instability by the number of major changes in the rate and/or direction of growth. Illustratively, we might quantify growth instability over a given time period by the degree of the polynomial trend equation which best fits the data (degrees of freedom considered). This is not to deny the difficulty of separating trend from cycle, to say nothing of the paucity of data. But when one recalls the central place of public investment in urban planning and administration (e.g., schools, streets, sewers) and that trend shifts are more likely to induce major amounts of in- or out-migration than cycle movements, the significance of growth stability is at least as great as cyclical stability in the urban context.

Again, city size is of interest. Industry growth trends, like seasonal but unlike cyclical patterns, are probably randomly distributed in time; industrial diversification tends to produce a mix of offsetting trends. Thus, the larger the urban economy, the more diversified its industrial structure, the stabler its growth, and the more it resembles the national economy in growth rate as well as growth stability. Of interest to the planner is the corollary that "nationalizing" the local rate of growth renders it more predictable.

Economic Growth and Development

The Demand Side

In the analysis and projection of regional economic growth and development, income elasticity of demand again comes into focus; with a steadily rising per capita income, time is on the side of areas producing income elastic goods. But in tying income elasticity to regional growth, we need to distinguish between the first quick response of buyers to a sudden (cyclical) change in income and a sustained, deliberate response to a slow, steady (secular) rise in income. Thus, where the durability (postponability) of a good is most relevant in a cyclical context, growth analysis forces us to think more in terms of the inferior-superior goods context of standard micro-economic demand theory -- from grits to beef and from bus to automobile transportation.

Beyond elasticity, shifts in taste patterns are elemental to the long view. The affluent society is characterized by a growing proportion of discretionary income and its reflection, discretionary spending. If, moreover, we are becoming ever more outer-directed and steadily trading local and regional identities for a more homogeneous national culture (via nation-wide television, Madison Avenue and ever easier migration), then the stage is set for capricious shifts in individual spending to amplify into massive tides of spending which engulf the whole country and ebb and flow as fads and fashions change. Manufacturing firms which sell in a national market find themselves alternately inundated with demand then left stranded on the

beach. The small (remote) urban economy built on a single product, or even a single brand, may be becoming much less viable in the new economy.

Satellite towns within the bounds of large metropolitan areas -- within labor commuting range of a large central city and/or a host of other small towns -- would, of course, be economically viable as part of an industrially diversified local labor market. Again, care should be taken to anticipate progress in transportation by generously delimiting the future commuting range -- by not classifying an urban area as remote (and vulnerable) too quickly. A construct analagous to the metropolitan area might be applied to less densely populated areas; employment services might well begin classifying and disseminating labor market information in the context of clusters of "independent" but economically inter-related small towns. It was argued above that growth stability but not cycle stability comes with city size and industrial diversification and it is growth stability that poses the critical problem for the small urban area. We need some careful quantification and statistical analysis of growth stability by city size and spatial position in the system of cities.

The Supply Side

Labor. The supply side of regional economic growth and development has long been recognized, even antedating the demand approach; a supply approach is implicit in the inevitable inventory of local resources which is the traditional beginning (and usually ending) point of the typical area development study. But the sophistication and analytical rigor that characterizes the demand analyses are rarely matched in the local resource-supply work. Occasionally the local labor force is placed somewhere in the spectrum of development between the shirt or work-shoe factory and the precision instrument factory. Seldom, however, is this dynamic conception accompanied by a sense of the time dimension of the industrial acculturation process -- do the steps upward in labor skill span a decade? a generation? What is the mechanism of skill transmission: through the industrial base and personal contact (from father to son) or through earnings and improved public services (from income to tax base to good schools)?

It was argued above that local specialization in unionized oligopolistic industries tends to raise local wage rates above the going rate for local skills, but the day of retribution arrives when local firms begin to decentralize; the new jobs needed to accommodate a growing local labor force must be supplied by new firms. But new firms may hesitate to enter an area in which the wage rate is high relative to labor skills; wage rates won from plush durable goods oligopolists may put the more competitive non-durables and export services out of reach in local area development efforts, at least until painful adjustments have taken place through protracted unemployment.

Entrepreneurship. While we all hasten to pay our formal respects to entrepreneurship -- inventiveness, venturesomeness, promotional artistry, organizational genius, and so forth -- as the critical ingredient in economic development, we usually hurry on to consort with more familiar acquaintances -- equilibrium rates of capital formation, capital-to-labor ratios and the like. Either we must tackle the heroic labor of quantifying entrepreneurship or we can write epic poetry on this subject, but there can be no satisfying regional growth theory without explicit and central recognition of the key factor in the capitalist system, and in long run growth in particular.

Chinitz has advanced the intriguing hypothesis: like father, like son; an area with many small businesses tends to spawn more new entrepreneurs than a corporate economy. To this let me add the concept of long waves in local entrepreneurship. A local inventor is successful and creates a new product or process; he probably innovates and begins manufacturing in his home town (he may hedge by retaining his old job temporarily, he can hire trusted assistants more easily at home, after Florence and Greenhut); the Cinderella firm pulls the local economy up with it; but success and large size bring bureaucracy (Parkinson's Law) and the emphasis shifts from creativity to efficiency and stability ("scientific menagement"); the maturing of the industry in which our firm is a leader will probably occur at about the time that the local firm is decentralizing operations to tap its nationwide market more efficiently; the combination of a maturing and decentralizing industry slows local employment growth below the rate of natural increase in the local labor force creating a local employment crisis; creativity is once more emphasized and perhaps more important financed and subsidized under frantic area development efforts -- the entrepreneur is king, the king is dead, long live the king. To this we may add the further condition that entrepreneurs tend to bunch in time and space as they attract each other -- to learn and steal from each other.

Turning to problems of quantification and data availability, the early returns from the manipulation of patent data, as a proxy for inventiveness, have not been exciting but with some imaginative taxonomic work we may yet gain some insights from this data. Specifically, progress needs to be made in the delicate business of linking patent classes to Census industries so that local patenting rates may be standardized for industry-mix, as patenting comes much more easily (nominally?) in some processes-industries (e.g., chemistry-chemicals).

Further, except for a passing glance from Ullman five years ago, this writer knows of no one who has experimented with the regional distribution of research and development associated persons. An urban-regional classification of natural scientists (from, say, American Men of Science), engineers (various Who is Who volumes) and others wants for nothing but labor power and/or automatic data processing. And even relative managerial skills may be compared between urban areas with imaginative surrogates, such as the degree of exposure of middle and top

management to post-graduate education, the speed with which bright young men are advanced, time lost through labor turnover and strikes, and so forth, prevailing circumstances considered of course.

Capital. Money capital is so spatially mobile that the existence of a local capital market must first be established. While large corporations probably do relate to the national capital market almost entirely and probably can be disregarded, very small local firms and especially prospective new entrants into business may well have to depend on neighboring lenders and investors. And these small fry are the stuff of which area development is made. Do urban areas differ notably in the cost and availability of debt and/or equity capital? fixed and/or working capital? Did the Detroit bankers who courageously financed the speculative infant automobile industry fifty years ago, when their eastern counterparts were hanging back hamstringing their local entrants in the automobile development and production race, also go through a "long wave" metamorphosis like their customers, growing more bureaucratic and conservative with success? Have Detroit bankers become so accustomed to the big, safe loan to the giant oligopoly that they have forgotten how to scrounge for new business and perhaps have no stomach for it, and will they respond to the challenge of the current industrial stagnation in Detroit by reverting to their old venturesomeness?

Turning to equity capital, is local wealth more inclined to sponsor local inventors and promoters than outside ones and, if so, do the local rich differ noticeable between urban areas? Are the Houston rich brash, aggressive gamblers and the Denver rich more staid and contented -- contented with low yields on geographically-diversified, high-grade bonds? Finally, do urban areas with greater income inequality (more very rich persons) provide better sources of local risk capital than more equalitarian urban economies. The quantification and data necessary to answer these questions would seem to be at least as dependent on hard work as on great insights or statistical artistry -- the bottleneck is as much research money as unusual talent. And the protection of confidential information in the sensitive financial realm -- the disclosure problem -- seems quite manageable, at least in the larger urban areas.

Land. Let me merely introduce the fourth classic factor of production, "land", with a direct quote from the manuscript:

> "Suppose automation reduces the force of the labor factor in industrial location by reducing the amount of direct labor input and unionization continues to spread its influence equalizing wages everywhere, especially for the critical skilled labor. Suppose differentials in capital supply to be of only minor importance with giant enterprise omni-present among industries and areas, creating thereby a national capital market. Entrepreneurship could then become the critical location factor as substantial inter-regional variations in inventiveness and venturesomeness remain. But what is more footloose, in aggregate, than an entrepreneurial complex; there is no obvious reason why an exciting and fruitful inventive-innovative environment might not be developed in a pleasant place to live rather than a less pleasant one. Inventors and scholars and promoters are human too -- at least their wives are. What could be more logical than for these intellectual-industrial centers to be consciously implanted in two distinctive environments: places which offer natural beauty and outdoor recreation, and places which offer the height of urban culture and indoor recreation? The case of Palo Alto and Santa Monica, California as centers of research and development does not need extended argument; it is hard to argue with success. Similarly, the superb consumer capital (e.g., museums, libraries, theatres and so forth) in New York and Boston have made it possible for those areas to hold their own in competition for research and development activity with the aforementioned garden spots."

The Matrix of Goals

"Opportunity Cost" at the Policy Level

The literature of urban economics is remarkably free of any discussion of the mutual compatibility or inter-relationships of these four goals, affluence, equity, stability and progress. To the extent that any particular couplet of goals is incompatible, then the pursuit of one goal has as its "opportunity cost" the sacrifice sustained by not emphasizing the other one instead.

Of course, the matrix of goal associations may reveal at least an equal number of positive correlations -- conjoined, consistent goals. This latter situation might seem, on first blush, to connote the best of all possible worlds-- to be able to have one's cake and eat it. But further reflection reminds one that, if the darker side of each picture still remains, a rash of positive correlations between these goals may mean the compounding of good fortune one place and ill fortune another. For example, if the rapidly growing areas are also the most stable, then the stagnating areas are doubly cursed by having heavy cyclical unemployment piled on top of heavy structural unemployment -- not too inaccurate a description of a distressingly large proportion of the local economies in Pennsylvania, West Virginia and Michigan.

Substitutability and Complementarity of Goals

Limitations of time and space prevent elaborate detailing of the probable interrelationships

between the level, distribution, stability and growth of income and employment, for there are twelve separate couplets to be explored (counting opposite directions of causation separately). Still, some feeling for the analytical power inherent in the concept of a matrix of goals can be gained from even a highly impressionistic sketch.

Level-Distribution. The level and distribution of income may be linked through the industry-mix, with the durable goods towns exhibiting both affluence and egalitarianism largely through the aegis of a third force, unionization. But if the impact of automation continues greater in manufacturing than in services, and the (income elastic?) services continue to grow faster, chronic unemployment will plague manufacturing areas. We may witness the evolution of an ambivalent income pattern in factory towns in which a relatively equal distribution of income among the upper four-fifths of the population is associated with an unusually impoverished lowest fifth. Thus relative high per capita income, relatively equally shared by the majority may be associated with a relatively large indigent class. The highly productive, automated-factory worker would be financially able to assume the heavy local welfare load; what is in question is whether "solidarity" is translatable into noblesse oblige.

Level-Stability. Again, working through the local industry-mix, a tenuous case can be made for the association of high income levels with cyclical instability. It is presumably well accepted that durable manufactures are more unstable than non-durables and services are the most stable of all. If wage rates run from high to low in the same order, then affluence and stability become substitutive goals and a trade-off is a bitter fact of life to be faced by the area development commission. Reversing the causation, stability operates to raise real income by eliminating alternating periods of excess capacity and capital shortages, economizing on urban capital.

Level-Growth. Traditional economic analysis has been more concerned with the income level-economic growth interrelationship than any of the others because relative wage rates (and job availabilities) are the key labor allocating mechanism. But if we face a decade of heavy structural unemployment, the interregional redistribution of labor may not proceed smoothly. While a job may induce an unemployed worker to move from a declining area, an increased probability of gaining employment may not: moving to an area with a lesser rate of unemployment does not provide a job. Depressed area legislation and Federally-extended unemployment compensation, however desirable, further complicate labor allocation. Forecasting local income levels and aggregate population change is increasingly an exercise in political economy.

Stability-Distribution. Apriori, one might argue either side of the income stability-distribution case, pointing up the need for empirical work. Clearly, a heavy local cycle implies severe unemployment, poverty and income inequality, but we can return again to the fact that the unstable durables are the most heavily unionized and hence most subject to that egalitarian influence. Thus a fluctuating aggregate local income would seem to be associated with a fluctuating coefficient of local income inequality. In prosperity the high income heavily-unionized durable goods areas may be relatively more egalitarian and in recessions relatively less so.

Growth-Stability. One feels intuitively that a rapid rate of aggregate growth should foster cyclical stability, if for no other reason than that it would swamp the local cycle, especially in an era of mild recessions. The verification of this impression is complicated by a subtle measurement problem. While the cycle in a rapidly growing local economy may be rather large when measured around the trend (with trend removed), a cycle trough that is barely lower than the preceding prosperity peak is a local "recession" only in a very special sense -- a reduced rate of in-migration with local employment holding steady is relatively painless (e.g., Los Angeles and San Diego in the 1954 recession).

But rapid rates of growth are hard to maintain; often they are built on the transitory stage of most rapid growth of a single industry or firm, and even if area growth spawns new industries (via growth of the local market, input-output linkages and so forth) transitional discontinuities are inevitable. In short, rapid growth probably almost ensures growth instability.

Growth-Distribution. Rapid aggregate growth serves to reduce cyclical and structural unemployment, reducing income inequality. A partial offset lies in the tendency of growing areas to attract the poor unskilled rural migrants, continually replenishing the local stock of low income residents. Rapid aggregate growth also contributes to intra-area, inter-governmental income inequality by creating suburbs and satellites (political fragmentation), politically separating the rich and the poor, and divorcing tax base from public service needs. This is reinforced by the tendency of the poor in-migrants to settle predominantly in the older (lower rent) central city.

Obviously, we might next move to a higher level of analysis and look for patterns combining three or more of the attributes (goals) and, ultimately, describe common syndromes by which urban economies might be classified. We have here also an unexcelled opportunity for the application of multiple correlation and regression analysis and other multi-variate analysis.

COMMENT ON: URBAN-REGIONAL ECONOMIC ANALYSIS: CONCEPTS, MEASUREMENT AND DATA*

Benjamin Chinitz, University of Pittsburgh

I have long been disturbed by the preoccupation of regional economics with the problem of allocating national economic activity among regions. The implicit assumption that the nation moves ahead under its own momentum with little or no reference to what is happening in different regions has always struck me as artificial and impractical. True, there is a sharp difference in degree between international mobility and interregional mobility. But the degree of interregional immobility is surely high enough to qualify as an important variable in national economic development.

If we view the national economy as independent of regional patterns we threaten the very raison d'etre of regional economics. Who cares whether City A is growing faster than City B if the sum of the two is a constant--unless you happen to own real estate in City A. If the growth in productivity in the nation is a given, then if I make City A more efficient I must somehow be making City B less efficient and I lose interest in the whole venture.

On the other hand, if I view national performance as an average of regional performances, I leave the door open for measures applied in particular areas which can raise the average and this offers a much more compelling motivation for the geographically impartial social scientist or public servant.

It is in this context that I find Wilbur Thompson's manuscript most refreshing and reassuring as to the future course of regional economics. For he has unequivocally taken the position that the urban area--like the time honored industry--is a meaningful unit of analysis, an arena in which to observe the processes of economic development. He recognizes full well that the urban area is very much of an open-ended economy and as such is vulnerable to seasonal, cyclical, and secular shocks originating outside the system. But he makes it quite clear that the tracing of exogenous impulses is not the be-all and end-all of regional economics.

Turning now to some specific points in Thompson's paper, I call attention first to his "roll out" concept in the theory of local wage level determination. He makes the point that high wages in the export sector make for high money income which may partly be dissipated in high prices for locally produced goods and services because the high wages in the export sector "roll out" to the local sectors. We in our study of the Pittsburgh economy find substantial evidence to support the "roll out" hypothesis but not exactly in the form implied by Thompson. We find indeed that the high wages of steel workers make for relatively high wages in other sectors of the economy but not in all of them, and especially not in local trade and service sectors.

How do we explain such a pattern? In the first place, demand factors, on balance, work against the "roll out" effect. Pittsburgh's export sector employs a lot of men but very few women. But men and women typically come in pairs so that the supply of women is a function of the demand for men, as much as it is a function of the wages paid to women. With an inelastic supply and a low demand, wages should be low. They can't be low in the export sector because this is the battleground for the countervailing' monopolists. So they are low in the trade and service sectors where union power is a lot weaker.

A second aspect of demand which works against the "roll out" effect is the generally slow growth of the export sector. The labor force is constantly growing through natural accretion. Outmigration, though large, is not large enough to prevent that growth. With little or no growth in the demand for labor in the export sector the influence of the export sector on the whole area's wage pattern diminishes.

In other words, the "roll out" can work through the demand side in a rapidly growing area whose export sector is not lopsided in terms of the labor force it employs. But if the demand route is blocked, then the "roll out" can only work through the exercise of monopoly power against the employers in local trades and services. The outcome depends then on the extent to which these industries are unionized. The construction industry fits Thompson's model very well, but most trade and service activities do not.

Thus, in our comparison of Pittsburgh's wage structure with that of 33 other areas, we found that Pittsburgh has one of the most 'stretched out' wage structures, occupying very high ranks in manufacturing and construction industries and very low ranks in retail and service industries.

Thompson, in his full-length manuscript, though not in his brief paper, develops a growth model of the urban area in which he deals very effectively with the implications of growth for the filling out of the local sector and the feedback effects of this on the export sector. There's no question that aggregate size is a

* Paper by Wilbur Thompson presented at 122nd Annual Meeting of the American Statistical Association, Minneapolis, Minnesota, September 9, 1962.

critical variable in determining the structure of the local sector. Self-sufficiency requires a minimum scale of demand. This scale varies between activities. Hence, as areas grow in size their complex of local activities changes.

But in my own work, some of which was stimulated by some worksheets I inherited from Thompson and Mattila, his colleague at Wayne, I have been impressed by the substantial variation around the size function. Recently, for example, I computed per capita employment in 127 trade and service industries in 67 urban areas and identified Pittsburgh's rank in each industry. Pittsburgh is in the first decile in absolute size, but in 94 of the 127 industries it was below the median, and only in 7 cases was it in the first decile of the distribution.

This leads me to my final point which is, that we need some new research technology in trying to explain the dynamics of the urban economy. So far I can only point to one distinctive tool, the mix test. This crops up again and again in almost every major work in the field. What would the area look like with U. S. weights and regional values? This has been applied to wage levels, income levels, rates of growth, cyclical behavior and so on.

This is certainly a useful exercise and a necessary one. I don't mean to demean it. In fact, in our work in Pittsburgh we have used it extensively to explain why the Pittsburgh economy behaves as it does.

But we need other new tools and approaches to cope with the residuals. Thompson refers repeatedly to multivariate regression analysis and has used it extensively in his own work. In principle it is hard to argue against the potential value of this tool. But the results are so often disappointing and frustrating, mainly, I think, because the tool is too blunt to deal with the subtle variation we are trying to explain, given the kind of data available to us. Our observations are too few, and our variables are too aggregative. Both problems could be alleviated if we had more data on plants instead of industries and on households instead of the labor force. With the aid of the computer, we can digest a lot more; hopefully we can convince the federal and local agencies to increase the feedings.

DISCUSSION

Charles M. Tiebout, University of Washington

Professor Thompson's paper, as he notes, is a condensation of a much larger study. In consequence, it is somewhat difficult for the discussant to play his traditional role of critic and knit picker. Further, rather than ready to pick, I find myself more in a "hear hear" role. The most appropriate thing to do, therefore, is to point up a couple of the implications, possibly extensions, of his analysis.

Thompson's discussion of local business cycles leads him to conclude that "elaborate efforts to measure the local multiplier do not rate a high priority." The simple multiplier formulation implicit in his discussions is of the following variety:

$$\Delta Y_T = \Delta Y_X \times \frac{1}{1 - \frac{\Delta Y_L}{\Delta Y_T}}$$

where Y_T = the total local income

Y_X = the autonomous local income derived from exports

Y_L = the income derived from production for local markets

The local consumption function component, $\frac{Y_L}{Y_T}$ assumes the marginal and average propensities are equal and constant--although this assumption is not critical.

Given this formulation, Thompson is correct noting that proportional changes in the level of exports produce proportional changes in the level of total activity. If one's interest is in comparing the cyclical sensitivity of two regions with this type of formulation in mind there is no sense in actually measuring the multiplier. Indeed, it would amount to measuring a tautology.

My concern here is twofold: (1) the relevant propensities may not be proportional, i.e. a 20 per cent decline in exports may not imply a 20 per cent drop in total activity; and (2) measurement of the local multiplier has many other uses aside from measurement of cyclical stability. The first point can be immediately appreciated when it is recalled that the local consumption propensity is really a combination of a number of propensities such as, consumers, business investment and housing investment.

One of the other uses of the multiplier is in forecasting long-run growth. The traditional approach is to estimate the changes in the level of exports and, via the multiplier, estimate the total change. While this technique, like any forecast, has pitfalls, it is still useful and requires a multiplier measurement.

In this regard, there is a difference in the use of the multiplier in measuring cyclical vis-a-vis growth impacts that is not sufficiently recognized. Cyclical imports relate to short-run changes. Within a region, per capita incomes rise and fall over the cycle. The appropriate propensities, lags aside, are marginal propensities. The availability of empirically determined marginal propensities at a regional level is all but nil. Long-run growth may prevent a different issue. For areas such as Los Angeles, Phoenix and Miami, the major change in total income comes about through increased population, not increased per capita income. If new residents' spending habits are the same as established residents', a cross section study of average propensities is relevant. Thus one of the empirical tasks of developing time services is not so urgent, and even a relatively simple multiplier is useful.

Thompson's points with respect to income distribution and local public finance are well taken. My only concern is that we do not confuse the issues. It is true that tax colonies exist within our metropolitan regions. It can also be asserted that some areas are "too poor" to afford "needed" services. Yet it is also necessary to keep distribution questions separated as much as possible from those of public goods and services provision.

Thompson's notes that "free local public services could come to be the principle instrument of income redistribution in the coming decade." This raises a troublesome question in the area of fiscal federalism, who is to bear the net tax-benefit loss in this redistribution? The argument is that the New York City well-to-do alone should not bear the total redistribution cost for the city's slum improvement. The suburbs in the metropolitan area should also be required to bear part of this cost. The notion is that they are also members of this community, even if the political boundaries are not coterminous with "community" boundaries.

Historically with less mobility and less rapid urban growth this position, it would seem, had more merit. Today it is not clear that the well-to-do along Philadelphia's Main Line or Detroit's North Shore should be solely responsible for the welfare of the in-migrants to these communities. This is a Federal function. Indeed, much of the redistribution in the form of goods and services has been federally subsidized and, I guess, for just this reason. Without such a subsidy the few well-to-do in an otherwise poor community will bear a disproportionate share of the costs compared to their equals in other communities. To argue, as some have, that the wealthy benefit, misses the whole point. If they do, then by any reasonable definition there has not been a redistribution. All this points out that we should be quite careful in looking at redistribution versus other governmental functions in metropolitan areas.

To repeat, I suspect these points are not new to Professor Thompson and only hope to extend his fine analysis.

DISCUSSION

Kirk R. Petshek, University of Wisconsin

Professor Thompson's paper -- and much more so the first five chapters of his book -- are a real contribution to urban economics, the almost non-existent science! One is most impressed with the breadth of his investigation and the imagination he brings to bear on each of the aspects he decides to discuss. While he fully recognizes the great amount of spade work yet to be done, his broad vista performs a much more needed service at this time than later detailed statistical investigation may contribute later; and I say this advisedly to a group of statisticians!

One looks forward with keen interest to the second part of the book. While it is vital that Professor Thompson's ideas are widely and extensively discussed, one feels intensely inadequate in trying to pick fly specks out of an excellent pepperpot soup!

I don't think it is Dr. Thompson's or the chairman's intention to keep the discussants to the brief paper presented; rather, I will refer to concepts and thoughts of the chapters themselves in trying to comment on some of the basic ones.

1. Migration

Professor Thompson, starting his analysis with the local labor market, is not fully satisfied with the entity of the Labor Market Area we all have used so long. He introduces the idea of the "extended urban family", which exceeds the commuting area. This concept tries imaginatively to explain the long-run development of the larger area by including those surrounding units which, in good times or bad, supply the additional labor needed or are respectively the recipients of urban outmigration. I am afraid the concept is more imaginative than useful. It is hard enough to try working with the "situs" problem, that is, distinguishing between an urban area's daytime and nighttime population: yet for many practical problems some way of distinguishing between them must be found. For instance, thanks to political fragmentation, it sometimes becomes necessary to determine the proper rate of unemployment for areas smaller than labor markets -- to the unending horror of statisticians in the federal government. In order to be able to base public policy decisions on this figure and report a percentage which is truly comparable to the unemployment rate of the total labor market, a relationship between unemployment, which is always measured by residence, and the resident labor force must be found. In other words, I feel that intra-urban statistical concepts would have been more important than supra-urban concepts.

But much more important than the difficulty of working with the idea of the "extended urban family" is the implication behind it, which is clearly that migration of labor is a frequent and automatic response to economic conditions. What is more, Dr. Thompson's belief in that probability permeates the work. True, he occasionally admits sluggishness and migration, or its lopsided character (this might best be expressed by calling it a tendency in favor of "to-migration" and at the expense of "from migration"). We all know that youth, or those entering the labor force, are apt to migrate if they find no suitable employment. But does this tendency really extend to those with families and roots? The assumption of migration is implicit in Dr. Thompson's statement that forced migration can only be avoided by holding population constant over time, which, while not accepted as a national goal, "is accepted as a prime community goal much more unequivocally" (Chapter 4). He continues that, if social costs are involved, it is not acceptable to force mass migration in order to gain an increase in per capita GNP! But the real question is, how do you force mass migration anyway, even if it were acceptable, unless it is in wartime? In the same vein, he speaks of a schedule of costs with different rates of forced migration, both in and out. Are these the costs to the total population involved as of today, or is it the cost to the community during the next decade? In fact, he holds information on this social cost factor necessary for public policy.

I submit that this concept of migration as a true safety valve is unrealistic. Not only is it hard to force people to migrate in peacetime, but they fail to do it in sufficient numbers in response to economic conditions, and public policy must not count on it. In fact, this unwillingness to migrate, even where it may be socially desirable, led to the introduction, in the Manpower Training Bill, of a clause providing for government payment for those who could not find employment locally. With a great political hoopla this clause was eliminated from the Bill. It seems more likely that the recipients of inmigration were more afraid than those who were about to lose population. Be that as it may, both the necessity of introducing such a clause and the political motivations of those eliminating it, seem to prove the point that we could not even achieve forced migration in peacetime, let alone take voluntary migration for granted.

While it is true, finally, that there is a slackening migration of nonwhites into areas with a larger-than-average rate of unemployment, it is equally true that there is little evidence of their outmigration as the economic conditions deteriorate.

2. Wage Roll-Out

The next logical step would be that, if migration does not work properly to adjust urban markets, an equalizing tendency within any one

labor market area will ensue -- Dr. Thompson calls it "intra-area wage roll-out". While this is a good concept, which clearly holds true for small and isolated urban areas, one can't help wondering how much it is possible to generalize for large metropolitan areas. This point has been made convincingly by Dr. Chinitz. Is the wage roll-out likely to happen in an area economy of less than full employment? While the larger wages of steel workers are apt to increase the cost of living throughout a steel-dominated area, would not the simultaneous existence of sizeable numbers of unskilled unemployed have the opposite effect? I am wondering whether Dr. Thompson's reasoning pays sufficient attention to this group, which today increasingly consists of non-white immigrants. Their influence might seriously impair the effectiveness of the wage roll-out.

In the same vein, I think that the effect of automation on the level of employment, income and labor competition should have been considered more closely (and that, I fear, is evident throughout his exposition). As it is precisely the high wages that bring about greater capital equipment and thus technological unemployment, the wage roll-out may not be as effective; actually, high wage industry and unemployed of both kinds just described may easily exist side by side in the same area.

Dr. Thompson argues that as local firms expand in new plants elsewhere, a local "depressed area" problem is created, as the high wage rates make it difficult to obtain other industries, even durable goods, while nondurable industry is snatched away by cheap labor availability elsewhere. The point I am trying to make is that, if other locational advantages exist, the wage situation is not likely to stand in the way of attracting new industry, as the existence of pools of surplus labor may make an area attractive, even if wages are high among the employed. Nor is it likely that the wage situation would be the reason why the large local firm expands it's facilities elsewhere, as these firms are usually organized by unions with industry-wide bargaining contracts.

Could it be that Pittsburgh and Detroit, which Dr. Thompson is quoting, lack opportunities for industries using women labor because their locational advantages cannot attract them, although pools of available women labor are large? Could it be that there is some unwillingness on the part of wives of high paid workers to get employed at low wages? And could the unwillingness of displaced workers to shift their occupations be also responsible? Finally, is there enough evidence that displaced workers in general are willing to shift occupations? Professor Miernyk of Northwestern University, as well as the experience of other textile towns, not only in New England, seem to contradict that. As long as they -- and even more importantly, their unions -- hope for a come-back, they are apt to resist getting into a new occupation at the low end of the seniority list, and the unions do not cherish losing members.

3. Urban Management and Efficiency

Throughout, the question of urban efficiency has, quite justly, a high priority rating. In fact, one wonders why some measure of efficiency (not necessarily maximizing it) does not appear within the matrix of goals Dr. Thompson sets up at the end of his paper: some interesting couplets would result. While public and private efficiency are not always kept clearly apart, one always hopes that public action would increase private efficiency. But more of that relationship in a moment.

In paralleling the poor supply of good public administrators with the stage in the firm's growth where it is limited by its inability to expand management efficiently, Dr. Thompson is joining the growing field of comparisons between public action and market behavior. While certain aspects of this parallel may be somewhat doubtful, I would like to point out that the limits of urban public management probably lie elsewhere (although we all agree how scarce good public servants are): the likelihood is that the limitation lies either in external factors (public facilities which are not expandable) or in the inability to induce private investment to follow public action, i.e. the investment in public facilities. In other words, unless the public servant can so plan public investment that private capital follows in the same direction, the public facilities may easily not be too effective as far as urban efficiency is concerned. Furthermore, quite often the public is unwilling to spend the amounts of money which are needed to replace obsolete public facilities. Proper planning for maximum efficiency of public investment and proper inducement of private capital can minimize idle capacity.

Transportation is a good example, which Dr. Thompson also uses. Are we willing to subsidize railroad and urban mass transit sufficiently to induce automobile owners to use it in preference to their cars by establishing frequency of service, new airconditioned cars, free parking facilities at terminals and easy transfer from one facility to another? We have much greater possibilities than we have been exploiting, if we are willing to fully use public inventiveness rather than parsimoniousness; it has been tried in some places.

The efficiency of downtown is another example, which only rarely can be accomplished by urban redevelopment, but should actually be accomplished by the business community involved. Clustering of "furriers, gourmet food shoppes and musical instrument stores in one area, and ten cent stores and work clothes in another" (Chapter 2) is probably a much more important device for center city than for outlying

communities, for which Dr. Thompson uses the example. In fact, there you may want to cluster furriers with each other and musical stores somewhere else in order to permit comparison shopping, in addition to clustering stores for the same income group in one area. Similarly, the need for face-to-face confrontation will segregate one kind of offices (general offices) from others (e.g. headquarters offices) etc.

Does Dr. Thompson's view that managerial competence in public service is a critical supply factor mean that only able civil servants can be inventive for the public interest or that, if they are, that they can carry out their ideas? Coming back to some of the above examples, downtown urban redevelopment is of no use whatever unless private investment is convinced of its usefulness and can be sold on the way public officials see the new downtown, or else its efficiency goes down the drain. Or take the transportation case: if the mass transportation company is unwilling to make the adjustment to frequency and convenience of their service which is needed to keep automobiles out of the central city, effective public control is necessary. But assume the city administration, which has seen the importance of this kind of action and has laid out a clear proposal of how the public agency can control a privately managed transportation company in the public interest, runs then into a chairman of the transportation company's board who is the main Republican fund raiser, while the chairman of its executive committee happens to be the Democratic fat cat -- how good is public managerial competence? (Let me assure you that this is by no means a fictional example!)

Leadership and public support are thus essential. Beyond that, the likelihood of good management in larger administrative units seems reasonable. Dr. Thompson believes that also, but he appears to believe that dis-economies of scale might fit in with regard to public management. This is not a convincing argument. If public facilities run into increasing per unit cost, as they well might, the starting of a new unit (e.g. a new filtering station) may be indicated. I fail to see any reason why the Viner cost curves should not describe the situation with regard to public facilities up to a point, while overhead is actually increasing very little. In fact, we have made a study in Philadelphia where we added on paper prototypes of relatively small (marginal) residential or non-residential developments respectively to the existing private capital stock in order to test the effect on public operating cost: we found that with regard to a good share of the costs of the various city departments, overhead costs did not increase substantially, and in some cases even variable costs were not affected.

This actually leads to what probably is the most important question regarding efficiency: should we measure it on an average or total basis, or should it be measured at the margin? It seems that cost benefit studies increasingly emphasize the latter -- Wheaton's speech before the AIP meeting a few years ago expressed it extremely well. I submit that, looked at marginally, additional population or industries or other development can often be added with negligible additional administrative cost. The points on urban efficiency which Dr. Thompson makes throughout his chapters could easily be clarified if he had clearly used the marginal approach.

Actually, the same thing can be said about the entire urban growth concept. If we take off from existing capital, both public and private, and talk about growth mostly in terms of added facilities, the entire chapter on urban growth might profit in its approach.

4. Effect of Government Activity

The performance of local government regarding the income structure is effectively shown as having an ever-increasing effect in the direction of greater income equalization, particularly where it is reinforced by federal grants-in-aid (urban renewal, mass transit, community health programs, etc.). Dr. Thompson's point which impressed me particularly is the fact that we keep speaking of tax progressiveness without considering the purpose to which local expenditure is being put; i.e. the influence on income distribution of the local fiscal action becomes much clearer by looking at both revenues and expenditures.

Less convincing is his paralleling of farm programs with urban welfare programs. Here I feel that farm support continuously increases the income inequality between large and small farms, while the family farm continues to be used as an emotional argument for farm support legislation. This is similar to the way in which widows and orphans are used as an argument against dividend and interest withholding, no matter how insignificant the effect on them!

There is a similarity here with the case for small business. Dr. Thompson suggests it as one of the possible solutions to unemployment, by advocating financial aid to unemployed factory workers for going into small business. It seems doubtful whether this is a viable long-term solution, except possibly in some of the world's grossly underdeveloped areas. The social reasons for supporting both the family farm and small business are there, but a coupling of the goals of personal independence and of efficiency would demonstrate the tremendous social cost of independence in business. It may be a better idea to endow leisure-time activities with the kind of content it had in ancient times where creativity could help the individual more than nominal independence in business.

5. Goals

This brings us back to the excellent chapter on the matrix of goals. Its importance lies less in the type of goals Dr. Thompson pairs up, than in drawing out attention specifically to this important point. Too often all of us (other social scientists as well as economists) imply the goals in what we are doing or take their clarity for granted.

In his paper, at the start of discussing the question of goals, Dr. Thompson mentions affluence, equity, and progress as some of the problems we ought to consider. However, in his further discussion, he actually does not discuss these broad goals. The chapter is most suggestive, but I feel it does not go far enough. While Dr. Thompson confined himself to the juxtaposition of only the various magnitudes he has discussed in other chapters, I feel that this does not go far enough, i.e. the real problems lie in the coupling of economic goals with other social goals, because this is really where the conflicts lie buried. The contrast between efficiency and democracy would seem one of the most important ones or, as mentioned above, the pairing of efficiency and personal independence. It would be extremely important to know the opportunity cost of one in terms of the other, and public policy decisions may be based on analyzing such goal pairs correctly.

This is not an academic game, but a very practical one. The vital problem is to show both scholars and politicians the kind of choices they have to make. The kind of experiment we carried out in Philadelphia might be of some interest. In asking a group of civic leaders to make choices between different categories of public investment, and indicate their relative importance and the time sequence each civic leader felt was most desirable for specific public facilities, we first presented them with a list of potential broad-gauged goals for the city as a whole (a city for pleasant living, an industrial city, a city easy to reach, an educational center, etc.). After they had made both choices, the kind of city they wanted, their suggestion about the importance and timing of individual public facilities or public investment, an attempt was made to show the leaders we had interviewed the consistency or lack thereof of the two choices made. We then asked them to see whether in the light of this discussion they wanted to reconsider one or other decision!

In a way, only this frame of mind makes us really useful to the policy-maker. The politician's normal compromise gives up one approach for the sake of another, but it is usually a temporary expedient without full realization of the actual opportunity cost. Do we prefer an area that satisfies the recreational desires of its people or one which draws a maximum of tax revenue from industry? What is our real indifference curve between the two? How much slum clearance are we willing to forego for the sake of attracting the community leadership into the central city to live? How much is active leadership worth to a public official in terms of official credit he receives for his accomplishments? How important is a stadium in terms of obtaining the support of the power structure, if it has to be built at the expense of a civic center? What is the relationship between public efficiency and private efficiency? That last couplet alone would lend itself to lengthy discussion.

Dr. Thompson's description of Flint, Michigan, as having high per capita income and high income equality, leads logically to the observation that industry there is mostly run by absentee owners. This in turn has a definite effect on community decision-making. The article by Dr. Robert Schulze (in Janowitz' "Community Power Structure") addresses itself to this very question of how local decision-making degenerates with absentee entrepeneurs. Are we willing to forego effective local leadership for high income per capita? The answer may be different from community to community, but the results should in each case become obvious to policy-makers.

Finally, in contrasting income distribution and stability, Dr. Thompson speaks of the building trades' annual earnings which might equal those of other workers, in spite of cyclical instability. However, should he not have considered the tendency by the elite among the workers to take as much work as possible for themselves and supplement it at the peak with cheap apprentices? In this connection, he asks whether high paid workers are willing to take the responsibility for the chronically unemployed manufacturing worker. The way he puts it, the answer is obviously in the negative. It seems to me, however, that he should have asked it differently: will the workers insist on a higher wage, if it is demonstrably clear to them that they will thereby cause increased unemployment? It happens that this discussant has done some research on this topic and has found that, in all the cases he could find, the union leaders backed down from their wage demand as soon as they could be convinced of the causal relationship.

Dr. Thompson has presented us with a most important analysis of the type of problems urban economists should solve. His analysis by factors, by income determinants, by demand and supply, etc. is extremely helpful, although as he keeps pointing out, much more work needs to be done -- and it should be done along the directions he indicates.

Throughout the work that ensues, however, we ought to keep in mind that the pieces need to be put together again, after the detailed work has been done, so that a living organism results. The economic base theory really does not explain the dynamics of an area's economy. Yet, an explanation of the likely change is not going to emerge unless

the synthesis which must follow the analysis takes into account other than economic considerations. The goals must be chosen, and the influence of government must be considered, in terms of those choices which are practically feasible. The synthesis we need must explain both political and economic forces which bring about metropolitan change.

Let us hope that the work Dr. Thompson has given us will enable us not only to do the work that he suggests, but also to keep our sights clearly on the political economy of metropolitan areas.

X

STATISTICS AND LEGISLATION

Chairman, Luther W. Stringham,
U. S. Department of Health, Education, and Welfare

Page

Statistics for Legislators - John W. Lehman, Joint Economic Committee, U. S. Congress..... 234

Statistics for Health, Education, and Welfare Legislation - Wilbur J. Cohen and Luther W. Stringham, U. S. Department of Health, Education, and Welfare..... 238

The Role of Statistics in Formulating the (Proposed) Food and Agriculture Act of 1962 - John A. Schnittker, U. S. Department of Agriculture..... 243

Discussion - Lloyd W. Woodruff, Minnesota Department of Taxation. 247

STATISTICS FOR LEGISLATORS

John W. Lehman, Joint Economic Committee, Congress of the United States

(The views expressed are those of the author and do not necessarily represent the views of the Joint Economic Committee, its subcommittees, or its members.)

One of the classic tools of business research is the market survey where the producer or distributor seeks to find out about the characteristics of the consumers of his product and the uses to which his product may be put. Without straining the analogy too much, let us begin by trying to understand a little about the average Member of Congress as a consumer of statistics.

Such surveys usually start with something about the occupational background of the prospective consumer. Over half of the membership in both Houses of Congress are lawyers. About a third of the Members have had some experience in business or finance. For the rest, there are medical doctors, journalists, engineers, teachers, and even ministers. A few members have quite impressive credentials in economics, as teachers or through government service. Senator Paul Douglas, of course, is a notable case. No Member, I believe, would identify himself as a statistician, although the economists, engineers, bankers, and perhaps others, will have some background in mathematics, statistics, and accounting.

Most Members of Congress who have had formal courses in economics or statistics probably took their work before the techniques and processes of aggregate analysis and the tools of national income, gross national product, or their components were available. The average age of the Senators and Members of the House would indicate that a large part of the Congress were in college before or about 1930, although of course there are many younger Members with more recent college experience. It is significant to recall that before 1930 only a small portion of the economic statistics now available were even in existence, let alone being published with any currency.

This does not necessarily mean that the average member has no familiarity with new measures and methods. Some of these statistics or techniques are in general public use, others he learns about through the kind of built-in seminar which is constantly provided by hearings on complex legislative matters. Fifteen years of the Employment Act with the annual economic reports of the President and the publications of the Joint Economic Committee have also contributed to many legislators' exposure to and knowledge of economic and statistical methods and language. But the picture of a Member of Congress as a consumer of statistics which emerges is still that of a generalist, with broad professional training and experience and, except for a very few, no specialization in economics or statistics.

Now what are the major purposes for which this Congressional consumer needs statistics? We start with the 13,000-14,000 bills and resolutions introduced each session of Congress and the 2300 which go through the committees, are reported to the floor, and passed. Not only are the measures increasing in number but also in complexity, and the use of numerical data in their presentation has become more and more essential.

As a committee member, the Senator or Representative must analyze and vote on some of these in committee. As an individual Member he must reach a decision as to his vote on the floor. He must be able to explain or justify his vote or position on pending measures to his constituents--in speeches, in individual discussion, or in response to correspondence. Above all, of course, he must satisfy himself that his part in the action taken is for the public good. He will have help from the executive agencies, committee staffs, and his own office, but in the end it is he alone who must find a basis for making decisions on literally thousands of legislative matters.

Statistics for legislators are, then, simply statistics for policy-makers. But these are special policy-makers indeed. They must make the hard decision more often, over a wider range of complex subject matter, and with a greater penalty for error than ever before in our history.

The time available for analysis and consideration prior to any policy decisions is unbelievably short. The need to draw a conclusion on some basis is unbelievably pressing.

Let me here pause to say that in considering statistics for legislative policy-makers I will draw my examples today from the area of economic stabilization. This is the area of legislative activity with which I have been most closely connected and--because of the intensive work of the Subcommittee on Economic Statistics--an area which we may be able to discuss with more precision.

This approach also leaves everything else wide open for my other colleagues on the panel. I dare say, however, we will all make a good many observations which have common application.

We start with a look at how the legislative consumer feels about statistics generally. One of the first recommendations of the Subcommittee on Economic Statistics when it was established in 1955 was to have included at some place in the President's budget document a summary of funds spent and budgeted for economic statistics. The thought behind this recommendation was to be able to make some kind of macro evaluation of the resources devoted to economic statistics--over time and in comparison with the changes of the economy as a whole. The story it tells over the years

since 1955 is one of a continuing desire on the part of the Congress and the Executive to provide adequate statistics for economic analysis and to improve our statistical resources. The attached table provides details on how this support has been distributed.

Even before the establishment of the Subcommittee on Economic Statistics, the Joint Economic Committee reflected the Congressional interest in statistics in its early reports. The first report by the committee staff was Current Gaps in Our Statistical Knowledge, issued in July 1948. The committee staff was also asked at that time to develop a publication to meet the growing needs of the Congress for current statistical information. The staff of the Council of Economic Advisers had a similar project underway for the Council. A cooperative arrangement was worked out which resulted in the publication of Economic Indicators. The popularity of Economic Indicators among the legislators resulted in a request to have it put on public sale and the Government Printing Office now has a subcription list of over 10,000 customers.

One result of the Congress having its own statistical publication has been to strengthen greatly the confidence in our regularly published statistics. The monthly booklet, Economic Indicators, not only presents a handy reference for the busy legislator but is a symbol of the complete bi-partisan interest of the Members of Congress in having the best set of facts available. Given the facts, they will interpret or combine them in the way which they consider appropriate to their point of view.

This is not to say that weaknesses in statistics might not be pointed out by any member of the Congress. But, with extremely rare exceptions these criticisms are on a professional, technical basis, and the type of question which any informed user might well raise. On the occasions when partisan implications have been made, a member of the critic's own party will often rise to the defense of the statisticians. An even more dramatic instance was the vigorous defense of employment and unemployment statistics against public attack last year. Members from both sides of the aisle rose to express their confidence in the integrity of the statistics and the statisticians producing them.

The legislator's first interest, then, is in statistics which are as technically correct as possible. His faith in the integrity of the figures produced by our statistical agencies, both public and private, is a continuous compliment to the statistics profession. But he will maintain this faith only as long as it can be demonstrated that the weaknesses and limitations which exist are inherent in the amount and kind of resources available, or the state of the arts, and do not come from any attempt to direct the results.

Given reliability and integrity of the basic series, what kind of statistics are apt to be most helpful to the legislator? Historically the answer has been those directed toward a single problem. In fact, many of the series we have today were developed to answer questions in connection with specific legislation. Our international trade statistics in connection with proposals for trade legislation, the Consumer Price Index in reviewing proposals for setting wage rates for shipyard workers, income statistics developed for the National Recovery Administration, and consideration of tax legislation, etc.

Today, examination of stabilization policies requires broader tools for the legislator, and, above all, statistics which portray and explain relationships. Here is where statistics and statisticians are increasingly making their greatest contribution to decision-making. Given various combinations of policy alternatives what are likely to be the results?

Ideally, at least from the standpoint of the harassed legislator, these relationships should be capable of being set down in the most summary fashion. For the Joint Economic Committee and the Council of Economic Advisers, this need for summary statistics has meant the development of the simplified economic model and the concept of the "potential" growth line. Even these devices have not been as well understood as they should be. And, of course, here is the hazard for the statistician and economist. The projections themselves can go wrong--or, what is more possible, too much is expected of them.

In some cases the attempt to reduce relationships goes even farther to what are virtually "rule-of-thumb" concepts. These are very popular indeed. For example, one rule-of-thumb frequently used is the assumption that the automatic stabilizers will offset about one-third of each $1 billion drop in Gross National Product. Other examples of the summary statistics are those that have been developed for revenue estimates in connection with the consideration of tax legislation: total revenue will decrease $3 billion for every $100 increase in the exemption, or each percentage point reduction in all individual income tax rates will mean a $2 billion loss in revenue--or the broader and more debatable measure, that each $1 billion of tax reduction will add $2.5 billion to Gross National Prodcut through the multiplier.

I am sure there are some in this audience who look very skeptically on this type of analysis. Let me simply say for them, that the alternative will be the calculation of similar figures by much less sophisticated persons, with much less knowledge of the limitations involved.

One way of examining the concerns of the Congress with the nation's statistics is to look at the objectives the staffs of the Council of Economic Advisers and the Economic Statistics Subcommittee have been encouraged to pursue in the biennial reviews of Economic Indicators. We start with the question of timeliness which is so familiar to all of us who have been on the production side of statistics.

A very real dilemma which faces most statis-

tics producers, in my experience, is whether to publish an inferior number and get it out sooner, or a better number which can only come out a day, a week, a month, later. The compromise is, of course, the "preliminary" figure, which is later revised. I suspect that forced to a choice, the legislator would pick the later but better figure and, as the Joint Economic Committee has done, ask that the whole process be speeded up. It is exceedingly disconcerting to a legislator to come out with a policy decision based on a set of "preliminary figures" only to learn a month or two later that the figures really went up instead of down.

More and more emphasis is now being put on seasonally adjusted series. The concept and usefulness of seasonal adjustment is well enough understood that it outweighs any possible confusion from having adjusted and unadjusted figures in the same table. Charts, particularly, use seasonally adjusted figures.

Wherever possible, more anticipatory series have been added to Economic Indicators. This is in response to a Subcommittee recommendation made after hearing from the Federal Reserve task force on anticipatory statistics. The interest of the Committee and other members of Congress in having figures on an earlier phase in the economic process rather than trying to force conclusions from preliminary data on a later phase is not new. The Dun and Bradstreet series on Business Men's Expectations was originally developed for the use of the Joint Economic Committee and first published in 1948. The Committee is also credited with doing much to stimulate the preparation and publication of the National Industrial Conference Board's survey of Preliminary Plans for Capital Spending.

Another question which comes up in connection with Economic Indicators is the broken series. That is, the revised series which is carried back a few years. I realize that there may be times when it simply is not possible to link series in any satisfactory way because of changes in concept, etc. But even in those cases, it is up to the statistician to give his best judgment on the relationship between the two series. He is in a better position than the user to make this judgment. The preference is to carry the series on the revised basis back to the beginning of the old series, or, at a minimum, to show for some overlap period what the numbers are on both bases.

So far, we have discussed existing statistics and largely in terms of their general use in connection with economic stabilization. In recent years a new dimension has been added in the form of policy questions about economic growth.

Here our tools are not as good. What is the return to society for investment in human capital? We know something of this by deducting the usual inputs from increases in output but it leaves us with a very large area of growth to account for. Or what is the return for investment in research and development? The Subcommittee on Economic Statistics has just completed a very illuminating set of hearings on "capacity"--illuminating from the standpoint of how little we really know about capacity and its relationship to economic growth. The hearings and report on Government Price Statistics and the included report of the Stigler Committee indicate that some of our use of price statistics strains them far beyond their original concepts. Here frontier research is needed on series some of which may be forty years old.

The statistician needs, then, to keep his present tools sharp and to continuously be on the lookout for new tools which are needed now or will be needed in the future to help the legislator shape policy. It is often said that a political democracy rests on an informed electorate and an informed government. In this last half of the twentieth century, in the kind of world we live in, statistics and statisticians will be looked to more and more to provide much of the information on which our private and public policymakers will rely.

Direct Obligations for Principal Current Federal Statistical Programs, by Agency, 1954-1963

(Fiscal years; millions of dollars)

AGENCY	1954 actual	1955 actual	1956 actual	1957 actual	1958 actual	1959 actual	1960 actual	1961 actual	1962 estimate	1963 estimate
Department of Agriculture:										
Agricultural Marketing Service										
Economic & Statistical analysis	0.9	1.0	1.4	1.4	1.6	1.6	1.6	--	--	--
Crop & Livestock estimates	3.7	4.0	4.6	5.1	5.7	6.2	6.4	--	--	--
Agricultural Research Service (production economics)	1.3	1.4	1.6	2.4	2.7	2.9	2.9	--	--	--
Economic Research Service (except foreign economic analysis)								8.3	8.5	9.3
Statistical Reporting Service								8.1	8.8	9.7
Department of Commerce:										
Bureau of the Census	6.8	6.3	7.3	7.4	8.2	8.6	8.2	9.6	10.8	13.0
Business & Defense Services Administration (construction statistics)	0.1	0.1	0.1	0.1	0.1	.2	--	--	--	--
Office of Business Economics	0.9	0.9	1.0	1.0	1.1	1.2	1.4	1.5	1.6	1.9
Department of Health, Education and Welfare:										
Office of Education										
Research Statistics	0.2	0.2	0.2	0.5	0.6	.7	.8	.9	1.0	1.3
Public Health Service										
National Office of Vital Statistics	1.3	1.3	1.4	1.4	1.6	1.6	--	--	--	--
Public Health Methods & Reports	0.5	0.5	0.5	1.2	1.9	2.0	--	--	--	--
National Health Statistics							3.0	4.0	4.5	5.2
Department of Labor:										
Bureau of Labor Statistics	5.5	5.4	6.6	7.1	7.5	8.0	10.5	11.1	12.4	15.3
Treasury Department:										
Internal Revenue Service										
Statistical reporting	1.6	2.0	1.9	1.8	2.4	2.6	2.6	3.1	3.2	3.9
Federal Trade Commission (financial reports)	0.1	0.1	0.3	0.3	0.2	.2	.2	.3	.3	.3
Securities & Exchange Commission										
Operational & Business statistics	0.2	0.1	0.2	0.2	0.2	.2	.2	.2	.2	.3
Total, current progrms	23.0	23.4	27.1	29.8	33.7	36.0	38.3	48.4	52.8	61.9
Periodic census programs	1.5	22.6	9.2	4.2	6.4	13.6	95.4	21.4	12.8	11.5
Total, principal statistical programs	24.5	46.0	36.3	34.0	40.1	49.6	133.7	69.9	65.6	73.4

Source: Budget of the United States Government, special analyses of principal statistical programs and Office of Statistical Standards, Bureau of the Budget. Details may not add to totals because of rounding.

STATISTICS FOR HEALTH, EDUCATION, AND WELFARE LEGISLATION

Wilbur J. Cohen and Luther W. Stringham
U. S. Department of Health, Education, and Welfare

One of the most striking features of American life today is the rapidity of change. All of us--in government, in academic life, and in business--sense the dramatic forces that are shaping the Nation of tomorrow.

• Our population is growing by 3 million persons a year--a quarter million every month. Over 65 million babies have arrived since World War II. Average life expectancy now exceeds 70 years, and we already have 17 million persons over 65.

• Major changes are taking place in where people live and in how they gain their livelihood. Metropolitan centers have been and will continue to be the primary beneficiaries of population growth.

• To an increasing degree the Nation's work in agriculture, industry, government, science, and the armed forces requires more and more the professional, the technician, the skilled--and less and less the uneducated and the unskilled. For these greater tasks, youth must be prepared and adults retrained.

• Changes are occurring in the things for which consumers are spending their money and in the prices which they pay. More is being spent for services and for quality; and in the merchandising world rapid obsolescence and a succession of new products are norms.

• Support for research and development has accelerated tremendously. Public and private funds for research in the past decade probably exceed the total of such expenditures during the entire previous history of the Republic.

• The relations between levels and units of government continue in flux, with new relationships being forged and old concepts being given hard reappraisal.

• Old ideals of opportunity, of welfare, and of security are being given new meaning by acts of Congress; and the rights of citizens and the prerogatives of government are being given new interpretations by the Courts.

• From outside our borders--and from space--we are confronted with the rapid technological advances of other nations and with the aspirations of the people of less developed lands.

These are some of the forces that dominate this time in history. They determine the manpower requirements, the needs for facilities, the kinds of services, and the avenues of future investigation. It is our understanding of these forces and of the issues and of problems which they generate that determines, in large measure, national policies and programs of action.

The legislative program of this Administration in the fields of health, education, and welfare reflects an appraisal of major needs of today and tomorrow. Legislative proposals in 1962 are focused on four major problem areas:

First, expanding and improving educational quality and opportunities;

Second, meeting the needs of our older population;

Third, defining new directions in social welfare; and

Fourth, improving and protecting the public's health.

In his State of the Union address, his Economic Report, and in a series of special messages on education, health, and welfare, the President has indicated areas of unmet need and has made specific proposals for attacking them.

To expand and improve educational quality and opportunities the Administration has recommended programs that would provide for:

• Construction of public school classrooms;

• Better teacher salaries;

• Improved teacher training;

• Loans to colleges for the construction of academic buildings;

• Federal scholarships; and

• The beginning of an attack to reduce adult illiteracy.

To better care for the health needs of the aged, the Administration has urged the enactment of health insurance under social security.

In providing new directions in social welfare Congress has already enacted legislation that stresses:

• Services instead of support;

• Rehabilitation instead of relief;

• Training for useful work instead of prolonged dependency; and

• More skilled personnel and more

efficient welfare administration.

To improve and protect the public health we are recommending legislation that will:

• Expand the number of physicians and dentists;

• Support the construction of new medical and dental schools;

• Strengthen the Federal role in support of medical research; and

• Plan and organize for comprehensive environmental health controls.

A very large number of facts, collected from many sources by many agencies, public and private, have contributed to our understanding of the economic and social needs of the people in the fields covered by these legislative proposals.

For example, the assessment of the need for health insurance for the aged under social security, is based on:

• The growing numbers of aged persons, the result partly of the longer life span.

• The rising proportion of deaths that are caused by the degenerative diseases.

• The steep upward trend of hospital costs.

• The limited and often inadequate income of most aged persons.

• The restricted coverage of older persons by voluntary health insurance.

We know that hospital stays are costly, that older people go to hospitals more frequently and stay longer. We know that prolonged illnesses in many cases entail expenses which exceed the total resources not only of the aged persons themselves but of their families as well. The belief in the feasibility of such insurance, furthermore, is based upon actuarial analysis that has yielded the proposed days of coverage, the deductibles, and the other features embodied in the legislation supported by the Administration.

Similarly, the educational proposals are based on statistics in considerable depth that have been evolved over a long period of time both within and outside the government. We have reasonably adequate figures on:

• The growing size of our school age population;

• Present enrollments and those that can be expected in the years to come;

• The educational attainment of the population and the proportion who drop out of school in successive years.

• The numbers of teachers and their compensation; and

• The numbers of full-time teachers who are not qualified by the standards of their own States.

Educational statistics, such as those of classroom shortages, have come in for a good deal of criticism, some of which probably is justified. Yet in fairness it should be said that the statistical community has contributed importantly to our understanding of this vital area of national concern. Unfortunately, educational policies have been at the vortex of strong feelings with regard to Federal responsibilities, States' rights, racial integration, and Church-State relations. Under these conditions it is difficult to assess the extent to which statistical shortcomings are the cause of our national failure to enact legislation that is in accord with the needs of the Nation.

The statistical support for our proposals in other areas could be cited. It is sufficient to say, however, that few serious legislative proposals, involving large numbers of people and affecting major areas of national concern, are ever formulated without prolonged and searching inquiry. Without a reasonable statistical base few of them are likely to emerge as law from our governmental process.

Statistics help to promote the appreciation both by Congress and the public of the need for action. In general, statistical presentation for this purpose must be simple and their pertinence to the issue under consideration must be apparent. Their impact is reduced by statistical qualifications and the use of words with specialized meanings that are not generally understood.

Here are some statements that meet these qualifications and which have had or are now having real usefulness in the consideration of important subject areas:

• 5,000,000 Americans are mentally retarded.

• Public assistance recipients total 7.5 million--4% of the population.

• Today 48 million people are enrolled in school; by 1970 enrollments will reach 60 million.

• There are 8 million people in the Nation who are classed as "functional illiterates"--that is, they have had less than 5 years of schooling.

• Of persons over 65 discharged from short-stay hospitals, only 30% have 3/4 or more of their bill covered by insurance.

• 2,000,000 disabled persons could benefit from vocational rehabilitation. Disabled persons rehabilitated into employment will pay, on the average, $7 in Federal income tax during the remainder of their work lives for each dollar invested in their rehabilitation.

• Hospital expenses per patient day averaged $9.39 in 1946. They are now about $35 per day.

• The average cost of medical education is more than $11,600. The cost of a dental education also averages more than $11,000.

• 43% of the 1959 medical school graduating class came from the 12 percent of American families having incomes of over $10,000.

• The educational level which a person attains is a product both of the education of the father and the family income. Where the father did not graduate from high school and the family income was less than $5,000 only 13 percent of the children had some college attendance. In contrast, where the father graduated from college and the family income was $10,000 or more, 89 percent of persons aged 16-24 years old had some college attendance.

These examples illustrate ways in which the statistical community serves the legislative process both in the formulation of proposals and in their presentation. Yet there is a question as to the proper relationship between the statistician and the official, particularly the policy-making official. This question involves the responsibilities of the statistician in relation to the political process. In discussing this relationship some historical perspective will be helpful.

The name statistics was first applied to matters important to the State. The term statistics is derived from the Latin ratio status and an equivalent Italian phrase. These phrases were coined in the later Middle Ages to designate the study of practical politics, or the statesman's art, as distinguished from the study of the history and philosophy of the State. To these words we owe the English word statist, signifying a person versed in public affairs and the word Statistik, which was coined in the mid-18th century as the German equivalent of ratio status.

In one of the earliest known definitions, statistics was defined as "the political science of the several countries." The earliest known use of the word statistics in the English language defines it as "the science that teaches us what is the political arrangement of all the modern states of the known world."

Both the Statistical Society of London and the American Statistical Association in their early days viewed statistics as "the ascertaining and bringing together of those facts which are calculated to illustrate the conditions and prospects of society."

Only a century ago statistics were defined in terms that gave as much weight to concepts of the state and of society as to the concepts of facts and figures.

In recent decades the definitions of statistics have given increased importance to number and the ideas of state and society have receded into the background. Today we often see statistics defined simply as the numerical study of groups or masses and their component units. Here are definitions taken from some recent textbooks:

> "The main purpose of statistics is to draw as reliable conclusions as possible from the results of experiments."
>
> Statistics "refers to a vast and growing body of techniques for collecting, analyzing, presenting, and extracting the meaning from statistical data."
>
> "Statistical methods are the techniques used to facilitate the interpretation of collections of quantititative or numerical data."

We believe that it is unfortunate to find these definitions of statistics that omit reference to the application of the results of statistical efforts. While we can pay tribute to the many accomplishments of the statistical community, we also should recognize that there is some failure, or at least a reluctance, to emphasize, so far as our social problems are concerned, the importance of applying the results of statistical investigations to the policy issues of our day. There are real shortcomings in the effectiveness with which statistical agencies serve the legislative process.

We are aware that government statisticians have various responsibilities. Many of them are part of the staff of an operating program. Their jobs are to portray program developments and to aid in effective administration.

Others are employed to provide the public with "general purpose" statistics on various aspects of national life. The basic act which created the Office of Education, for example, charged it with "collecting such statistics and facts as shall show the condition and progress of education." Other agencies are responsible for collecting and issuing data that will aid the business community.

Only a small proportion of the government's statisticians are employed specifically to aid officials who are responsible for major policy decisions, including the formulation of legislative proposals. There are important exceptions, of course. The Division of the Actuary in the Social Security Administration for many years has provided the Executive and Legislative Branches with computations of the program implications of policy alternatives relating to social security payroll contributions and benefits that can be provided with given income levels.

In general, however, "statistics for legislation" is mainly a question of trying to make do with what is available. We hear about the "information explosion" and the enormously expanded quantity of data that is pouring out of the new data processing machines. In our experience this so-called "information explosion" has not as yet had a corresponding impact on the production of those compelling statistics that are of such importance to the legislative process.

Perhaps the "right" statistics are being produced, or could be produced, but still not enough attention is being given to the study of the outpourings of data to ascertain the policy implications. To extract from the masses of raw data information that can be used in a telling way in the solution of policy problems, still requires the old-fashioned use of eye-sight, reading time, and thought. There is still no substitute for human intellect and judgment. Greater quantities of statistics, in fact, extend the need for the analysis that can be provided only by people and their brains.

There are major problems in getting statisticians to think about national issues and the meaning of the data they are producing. There seems to be too much defensiveness in some of the statements we hear rather too frequently.

Here is the first one:

> "Our responsibility is to the integrity of our statistics. We simply cannot become involved in their policy implications."

This is what might be called defensive position #1. It is based on the implicit assumption that statistics and policy do not make virtuous bedfellows. It reflects a longing for the "splendid isolation" in which statisticians speak only to statisticians.

We realize that this statement represents an honest effort to keep statistics from being the handmaiden of partisan political expediency. It represents a desire to protect the statistician in those situations where statistics are but one of many elements needed in a policy decision.

Nevertheless, we regard the shying away from any concern for policy implications as unfortunate. We can sympathize with what is basically a plea for a tranquil life. However, we do not accept the notion that a statistician will lose his virtue if he looks a policy issue in the face. He can and should be keenly aware of the issues of the day and continually ask "What data do I have or what data can I provide that bears, honestly and objectively, on such and such an issue?" The hotter the issue, the more emotional and partisan the dispute, the greater is the need for the unvarnished objectivity of the statistician who can keep his head out of the sand.

Here is another one you may have heard:

> "If you will just tell us exactly what it is you want, we will try to help you."

This we might call defensive position #2. It is encountered not so often by the top policy official himself as by his staff representative who is sent out to round up any facts that might bear on some question under study.

The demand by the statistician for an exposition of "exactly what is wanted" may have a pretty devastating effect on the staff man, sent out on the fishing expedition, depending on his personality, experience, and tenacity.

Actually he often doesn't know just what he wants or what would be most useful. He doesn't know what might be available, and he usually is not versed in the statistics-lore and lingo of the particular Section or Division to which he has been sent.

In all fairness it should be said that, if the staff man does not wither completely under the initial blast, he usually gets something of use. He will quickly learn, however, that the empathy of statisticians to the problems of the policy official range from those with considerable sensitivity to those who have little or none.

Now for defensive position #3:

> "This series is now being revised. If you will come back in about two years we can give you something that will be much better."

This position says, in effect, that no matter what statistics are now available, and no matter what time, effort, and resources have gone into their production, they simply are not worth using. This position suggests that the policy user would do better just to go away and come back at some indefinite time in the future.

A variation of this "come back later" idea is expressed as follows:

> "The last time that survey was made was about 10 years ago. A new study is underway, but will not be ready for several months."

The injunction to "go away and come again some other day" is particularly discouraging in the context of the legislative process. The legislative calendar ordinarily does not have that much flexibility.

A message, or testimony, or a draft bill usually cannot wait. If we are not ready, we simply default. We have to use what is available, often what is immediately at hand. We can't ordinarily wait on the revision of a statistical series or even on the orderly completion of data in process.

From the policy standpoint, therefore, it is of the utmost importance that the greatest possible attention be given to the issuance at least of initial tabulations and summaries of findings. It is a pity not to have major conclusions available until every last appendix and every last cross reference have been completed and the study has found its tedious way to final publication. The extra effort to make results promptly available means the difference in many cases between its use and non-use for policy purposes. It also is discouraging to find that quite elaborate studies sometimes are published with no summary of findings and conclusions. Sometimes it almost appears that the originators and analysts are conspiring to make the study difficult to use for policy purposes.

Because of the lag in the analysis and publication of data it is especially important that increased attention be given to the preparation of current estimates and projections into the future. Such estimates are the best guesses of experts who are familiar with all of the related basic statistics. Current estimates and projections enable the policy official to talk more persuasively about needs and problems than if he could only make reference to a base period several years in the past.

In concluding this discussion of people in relation to the channeling of data for decision-making we would mention the key role that should be played by Directors of statistical organizations within the agencies of a Department. These people are in a pivotal position. They have under their supervision a variety of statistical and research personnel. They play an important part in the planning of statistical programs and operations. They also have access to officials in key positions, often to the very top of the organization. They are in a particularly favorable position to know about and to sense current and upcoming needs. They can, therefore, perform a vital function of bringing information to bear on the legislative process.

How well research directors perform this dual function determines whether statistical information enters the policy formation process in a meaningful flow or merely trickles up in fits and starts through devious and doubtful channels. Policy officials should be able to regard the Directors of major statistical programs, not just as overseers of statistical operations alone, but as persons who understand their obligations in synthesizing the subject matter in a given field and their role as resource persons in the process of policy formation.

A strengthening of the role of key research and statistical personnel will help in the support of our legislative activities. Yet, it is only fair to say that the ways of legislation are intricate, and the forces bearing on an issue are not easy to measure. We must always realize that all the data in the world are of no use, if the climate of receptivity is unfavorable. Statistics alone can never insure victory, especially when they collide with emotion and prejudice.

Yet we can be sure of the long-term power of honest facts honestly obtained. And that is our joint responsibility.

THE ROLE OF STATISTICS IN FORMULATING THE (PROPOSED) FOOD AND AGRICULTURE ACT OF 1962

John A. Schnittker, United States Department of Agriculture

The proposed Food and Agriculture Act of 1962 would authorize supply-management programs for feed grains, wheat, and milk, the commodities for which the most serious surplus problems exist, and would modify or expand certain soil conservation, credit, and rural economic development programs of the Department of Agriculture. It was sent to Congress by the President in late January this year, was generally approved by the Senate, disapproved by the House, and is now in Conference Committee of both Houses, in a much modified form. Both the formulation and the administration of the two kinds of programs pose major statistical problems.

Commodity Programs

1. Feed Grains.

The proposed program for feed grains would authorize the use of marketing quotas and acreage allotments as devices for limiting production and marketings. A national marketing quota for feed grains would be proclaimed each year, equal to the amount needed for livestock feed, human food, seed, industrial uses and exports, less estimated imports and if Government stocks are excessive, less an amount to permit stock reduction.

The national marketing quota would treat the four feed grains -- corn, grain sorghums, oats and barley -- as one commodity. It would be translated into a national acreage allotment on the basis of expected yields. This national acreage allotment would be apportioned to States, counties, and farms on the basis of the average acreage of feed grains produced during the base period.

Each feed grain producer subject to the quota would be required to remove from production an acreage equal to the difference between his farm's base acreage and the acreage allotment for the year. Land diverted from feed grains would be put into conservation uses, and possibly would be used for forage after a few years.

The Secretary would be authorized to make limited payments to producers who divert feed grain land to conservation uses for three years.

Prices of feed grains would be supported at between 65 and 90 percent of the parity price as determined by the Secretary. Price supports would be available only to producers who stay within their acreage allotments and comply with the land use requirements.

Production on excess acreage of feed grains would be subject to marketing penalty at the rate of 65 percent of the parity price per bushel when marketing quotas were in effect.

The marketing quota would be subject to a referendum and would become effective only if approved by two-thirds or more of the producers voting in the referendum. If the marketing quota were disapproved by producers, feed grains would be without price support.

2. Wheat.

The minimum national acreage allotment for wheat which has been 55 million acres since the late 1930's would be eliminated, and would be replaced by an annual allotment based on the year's marketing quota, as determined by the requirements for wheat and the need to reduce wheat stocks held by CCC.

The Secretary would proclaim a national wheat marketing quota each year equal to the year's total needs for wheat for human food, seed, and exports, less estimated imports and so long as CCC stocks are excessive less an amount to allow for stock reduction. This marketing quota would be translated into a national acreage allotment on the basis of expected yields. The national allotment would be apportioned to States and counties on the basis of the average acreage of wheat produced during the past five years.

Each wheat producer would be required to put wheat acreage into conservation uses in proportion to the amount by which the new national acreage allotment is reduced below 55 million acres. Land diverted from wheat would be subject to conditions described above for feed grains. The Secretary could make payments on wheat allotment land diverted to conservation uses for three years and could permit producers to divert additional land from wheat up to 20 percent of the wheat allotment and make payments for such diversion.

As in the case of other marketing quotas, the wheat program would be subject to a referendum of the producers and would become effective only if approved by two-thirds or more of the producers voting in the referendum.

3. Milk.

The proposed program for milk and milk products was designed to achieve progressive improvement in dairy farm income, while reducing Government program costs to a desirable minimum, by introducing two new features into the dairy price support program: (1) A supply-management program with price supports provided at least two-thirds of the producers voting in a referendum approved it; and (2) A limit of $300 million per year on the Government spending to acquire dairy products to be used for domestic welfare and foreign assistance, plus regular expenditures for Special Milk and School Lunch Programs. Marketing fees on over-allotment milk sales would, if marketing quotas were in effect, supplement the price support operation.

This program was designed to require the producers of excess milk to bear directly the cost of acquiring and disposing of dairy products in excess of what could be sold or used effectively in Government programs.

So much for a birds-eye view of the proposed program. It should be obvious that programs such as this require a full kit of statistical tools. Two types of statistical information deserve special emphasis. The first type are the economic and statistical data and relationships that underlie the kind of estimates and projections that are needed to permit informed and intelligent policy decisions. The second type are the statistics required to administer the program itself.

The first group is perhaps more important than the second, since program design can do much to ease later administrative problems. To turn to the feed grain program for the moment, the national marketing quota would require estimates and projections of the total needs for livestock feed, human food, seed, industrial uses and exports for a marketing year beginning about a year after plans for the crop are made. By far the greatest of these uses is livestock feed. The purpose of livestock production is not only to satisfy consumers, but more than incidentally, to produce income for farm people. It follows that projecting the quantities of feed grains needed for livestock feed cannot be done in a price vacuum, but must be done in the context of prices that will return reasonable incomes to farmers and yet will be fair to consumers. This requires projections of complex relationships among the prices and consumption of the various livestock items. Ideally, this would mean a complete model of the demand and supply relationships in the complex feed-livestock segment of our agricultural economy.

A considerable amount of research has been done over the years, particularly on the demand side. Brandow's work reported in Interrelations Among Demands for Farm Products and Implications for Control of Market Supply, Bulletin 680, Pennsylvania State University, is a good illustration of this.

Considerably less useful research results are available on the supply side. This is unfortunate since under present day economic conditions, the supply side of the price-making equation for farm products begins to over-shadow the demand side

Two rather deceptively simple statistical measures are of almost crucial importance in supply-management programs for feed grains. These are (1) yield per acre and (2) grain consumption per animal unit. Different combinations of these two can lead to evaluations that run the gamut from pessimistic to optimistic projections of the probable feed-livestock and price support situation.

There has been a strong upward trend yield of feed grains per acre. The ability to make accurate projections of expected yields is essential to any program which attempts to adjust feed grain production by limiting the number of acres used. Thus, accurate evaluations of the extent and the effects of fertilizer use and of the contribution of several other productive factors to the upward trend in yields in the past and their potential for future boosts in yields are needed.

Rates of feed consumption per animal unit have also shown a strong upward trend in recent years. This has helped to offset the embarrassment that might otherwise have accompanied the results of the sharp increase in yields, but it has not avoided distressing increases in surplus stocks. Here again the ability to measure the influence of the factors contributing to this upward trend and forecast their future course are highly important.

Here are further examples. We export substantial quantities of wheat, and this raises a host of additional questions involving evaluation of foreign statistics, foreign economic policies, and international relations. The fact that the farm production season in a year or more in length in some cases, and that the 1964 wheat acreage allotment, for example, should ideally be announced more than a year before any 1964 wheat will be marketed, and long before the foreign demand prospects can be adequately appraised raises further problems. An unexpected drop in the consumption of milk in 1961 complicates the problems in this field, and the fact that dairy animals produce both milk and meat is always

a potential source of statistical error in milk production or cow numbers.

The administration of supply-management programs requires additional detailed statistical information. A feed grain acreage diversion program requires information on acreage, yield, and production of feed grains on individual farms. Operation of such programs helps to produce statistical information that would not otherwise be available.

The Statistical Reporting Service is moving forward in the development of a program of enumerative surveys which are designed to improve the basic estimates of acreage, yield, and production of crops and numbers and production of livestock. A brief discussion of this program is being given before another session of these meetings.

Resources and People

When we come to the problems of land, water, and people, we have serious gaps in our statistical information and in analyzing such information. Much is known about our land resources in terms of land use, land classes, according to physical attributes,etc. We have to a considerable extent, information about land resources by sizes and types of farms and for broad regions. From the policy standpoint, data on land and its use are relatively adequate for policy formulation; for example, broad conservation policy. Statistics on water resources are probably less adequate. Perhaps in the West, where water has always been a critical resource, we know more about the problem than in any other area. However, in much of the central and eastern portions of the nation, where supplemental irrigation has become so important in the last decade or two, there is a serious lack of information significant for overall policy formulation. Variations among States with regard to water law and water rights makes it more imperative that we obtain a backlog of information on this resource and its use.

It is significant that in recent years that the problems of people in agriculture and rural areas are receiving more attention in policy-making. The statistical information to support realistic basic solutions to this problem have been rather weak, although actions are being taken to remedy this. In the past, statistical information on agriculture has generally dealt with farmers and farming as a whole without too much importance being attached to the dual complex of the agricultural population, one being the highly commercialized sector, and the other the low production and low income sector. The latter includes nearly twice as many people as the former, but it produces only between 10 and 15 percent of the total agricultural output. It is in dealing with the people in agriculture and their potential both in and out of agriculture that there is the greatest need for additional statistical information and particularly for analysis of information.

There has been and continues to be a real effort to identify and classify low income problems, and the characteristics of low income groups among farm and rural people. It is fully recognized that price and commodity programs, and even technical assistance programs built around agricultural production, can and will have little effect on the incomes and well-being of the low income segment of the population. Policy to deal with opportunity and income development is essentially general economic policy. It must have not only a substantial basis in the potentials and possibilities of people, but also in the potentials and possibilities of the national economy to grow, and of communities and areas themselves to participate in that growth.

In considering the income problem, considerable strides have been made in recent years in obtaining and analyzing some of the needed information. For example, incomes of farm, rural-nonfarm and urban families are now available, and intersector and interregional comparisons can be made. Significant measurements are possible which will relate to the economic development potential of populations and of areas to current economic development trends. With some detailed analysis, it will now be possible to measure the magnitude of employment and income opportunities needed for rural people and for significant adjustments that need to be made as between the farm and nonfarm sectors.

One of the most meaningful types of information needed, not only for agriculture, but economic growth and development in general, is a realistic measurement of economic development potential of areas. This is a tremendous job of statistical analysis insofar as agriculture is concerned, relating that sector of the economy to other sectors, analyzing mass data on incomes, occupational change, and input-output relationships. Much of the needed broad data are available for this on a State, regional and national basis, but there is also serious need for this type of information at a regional or area level. Statistical relationships of data, say at the level of 13 Southern States, are quite revealing and significant but they do not reveal much for an 18-county area, say of north-central Georgia. Some area studies which the Department of Agriculture has made helps us in this respect, but in these we have looked at only the resources of the rural open-country area.

In the Department of Agriculture, there has been and must continue to be a recognition that the effectiveness of much agricultural policy will depend upon nonagricultural facets of growth and development. Some examples of significant statistical analysis in process in this area in the Department of Agriculture follow:

1. The Rural Development Branch of the Economic Research Service has a contract for tabulations of unpublished 1960 Census of Population data for counties that will provide detailed information relating family and person characteristics and their incomes and economic activities.

2. A Special Project is underway in cooperation with The Agriculture Division, Bureau of the Census, in which the 1960 Sample Census of Agriculture questionnaires (and their 1959 counterparts) are matched with the 1960 Census of Population questionnaires for the farm households. The analysis of these data will permit the association of human and physical resource characteristics and utilization in agriculture for types and sizes of farms.

3. Work is underway to develop the data necessary for delineation of viable economic development areas that will encompass counties with concentrations of low income rural people.

4. Data for counties depicting the relationships over time of farm and non-farm family incomes are being developed and analyzed.

5. Data are being developed and refined that will permit the determination of the proportions of non-farm family and person and incomes which are required in order for farm people of comparable income earning capacity to be comparably rewarded.

Discussion

Lloyd W. Woodruff, Director of Research and Planning,
Department of Taxation, State of Minnesota

Resume'

1. Mr. Stringham suggested making a few observations about state and local government and statistics. If Minnesota is a reasonable example, the tremendous contrast between federal and state employment of statisticians almost defies comprehension. As an indicator of the contrast, total state employment in Minnesota approximates 15,000 while one major federal agency, the Department of Agriculture alone, exceeds 80,000. The Minnesota Taxation Department employs 550; the Internal Revenue Service, somewhere in the area of 45,000. In all of Minnesota state service hardly more than one hundred employes qualify as research personnel and a bare handful as true statisticans which I would define in this context as those working above the arithmetic, tabular level of computations. In short, the State of Minnesota does not employ many "high-powered" statisticians.

2. In search for a common denominator or an area of difference among these three papers, I found a lead to the perennial issue of the role of the expert (the statistician in this instance) in the political process. Picture a continuum of viewpoints about this role: at one extreme is the "three-branch" theory of our government where policy formulation rests almost entirely with the legislative branch; at the other extreme, lacking a short-title, is the belief in the inseparability of policy from administration--that even the lowliest of clerks engages in policy formulation as well as execution or administration. Accepting this continuum, we can locate Mssrs. Lehman and Stringham fairly close to the two extremes. Mr. Lehman, strange as it may seem since he is talking about a committee of the legislative branch, shows how statistics and statisticians can remain aloof from the policy process; the Joint Economic Committee, by which he is employed, functions as a research agency and initiates no policy proposals. Its staff concentrates on exploring basic problems, feeding its data into the legislative process but not attempting to obtain specific results. Mr. Stringham, though not centering his paper entirely on this thesis, does contend that statisticians should keep an eye on the political process--that they should be willing to face a policy matter, and I infer, to develop specific analyses needed in support of it. Thus, we have two roles for the statisticians in public policy formulation and enactment. Attempting to determine the "proper" role extends beyond the scope of this discussant's function--but what should it be?

XI

STATISTICS ON COLOR OR RACE

Chairman, Philip M. Hauser, University of Chicago

Page

Panel Discussion:

Warren M. Banner, National Urban League..... 250

Henry Lee Moon, National Association for the Advancement of Colored People..... 252

Henry D. Sheldon, U. S. Bureau of the Census..... 254

STATISTICS ON COLOR OR RACE

Warren M. Banner, National Urban League

The United States Government classifies its population into two major groups - white and nonwhite. Some of the officials who have been party to this farce are from our best educational institutions where even their own colleagues disagree with their simplification of so difficult an area of study.

History relates that hours on end have been consumed trying to decide what constituted an octoroon, quadroon, colored person, Negro, etc.

I do not say that knowledge of the number of people in any group is of no value; my contention is that Government, which must treat all of its citizens alike, must not arbitrarily divide the people into black and white, nonwhite and white, Negro and white, etc. I have heard some of the arguments made by the medical profession. I also have heard of the strong feeling in organized medicine that people must not be treated in groups. I know of the arguments made by law enforcement officers; and I also know that in instances where racial identifications were eliminated these same officers were no less efficient.

Groups have worked for years to get the press to eliminate racial identity in news items. This has been done in some places. Where it has been done the newspaper business has not suffered, and certainly the possibility for whipping up hysteria has been lessened.

In New York we are told that the schools (from the record) do not know the racial identity of those in attendance. However, a few weeks ago a report was released on school population showing the white and nonwhite enrollment. This report was made to prove that discrimination based on race does not exist. If the statistical count was a spot check, because it was needed for a situation which arose, I would not object. However, after the avowed purpose of the moment was accomplished, there should be return to the stated policy (which some of us know is not carried out in practice).

Confidential reports made by both public and voluntary agencies (white) seldom, if ever, make kind remarks about nonwhites, if the population has been so divided. This is even true in reports prepared for and by religious groups.

Voluntary agencies constituted of and working among nonwhite groups need to have general indication of the numbers they are concerned about but it really does not matter whether they find the number in a given situation to be 200,000 or 250,000 - 18,000,000 or 19,000,000. Whatever the count there are more people than are going to become "equal" in your lifetime and mine. Where we have reviewed budgets of governments and various subdivisions thereof, we have been able to get the politicians to be concerned about numbers and representation on public payroll only when it was a matter of life and death for themselves.

The recent articles on population explosion express grave concern about increase among those who "don't belong". Let's be honest. Few men accept the fellow who is unlike those in his immediate family. In the United States we have not taught our children the importance of oneness and the spirit that makes a nation strong because there is brotherhood under God.

Will the researcher be handicapped in his need for statistics in his effort to work among people for their development if Government does not continue to divide the population into white and nonwhite groups? You know that there are groups which count their own people, but which will not permit anyone else to do so. It has been pointed out that if the present arrangement is abandoned the nonwhite group would have difficulty finding out its numerical strength. Be assured that where, if, and when the nonwhite group needs to know its strength it can find out and in some instances the count will be more accurate than appears in the census. Much will have to be done before anyone who knows what is going on will believe that the statistics which are gathered on nonwhites are being used, even by Government itself, disinterestedly. Immigration quotas and the treatment of refugees (German, Hungarian, Congo, etc.) give you food for thought here.

Those who are interested in academic exercises will, of course, be unable to pull down a volume from the shelf and make some of their presentations. However, their demographic sketches will, in my opinion, be just as accurate without knowing the "identity" of about one-tenth of the population; to say nothing of the fact that even with the present statistical breakdown they have only generalizations.

I am unaware of the contribution that is made in the field of race relations without going among people and working with them. If you know of any place in the United States where the colored group does not live in "pockets" or special areas please enlighten me. This did not just happen. The ramifications of it are too familiar to you. From census tract and block statistics anyone who spends a few minutes in any of our large urban centers can tell the size of the nonwhite group. The same is almost equally true in rural areas. Both areas will remain the same for years to come. No major force is interested in any other arrangement.

As a member of the Council of Population and Housing Census Users for the 1960 U. S. Census, I am aware of the discussion held in connection with the preparation of the field schedules. Question was raised about the inclusion of religion on the general population schedule here, and elsewhere, and because of opposition to it this was eliminated. I had reservations about inclusion of "race" and wished this brought up. Unfortunately, I could not attend the meetings and an alternate went in my stead. He was to present my views on the question of "race" but did not do so.

Later I sat as a member of the planning committee for the study of social workers in 1960 - a 10-year report done by the Government in cooperation with the National Social Welfare Assembly and the National Association of Social Workers. Many social workers refused to answer the item on "race" in the earlier study and it was agreed that it had little significance for the new one. This item was eliminated.

Just a few years back, Senator James Watson, of the New York State Legislature, was able to have "race" eliminated from drivers licenses. A few of us had been working on this for about 15 years. Similar action has been taken in New Jersey, Washington, and possibly other states. The New York State Commission Against Discrimination has in writing the fact that since race was prohibited on civil service applications, state officials were calling the Motor Vehicle Bureau to determine applicants' identities. Part of the unfinished business of the previous chairman of the Commission was the elimination of race from applications for marriage licenses. Recently, in New York City, "race" was eliminated from birth records available to the public, and from applications for licenses to operate cabarets, public dance halls, catering establishments and coffee houses and their employees. New Jersey has also eliminated "race" from birth and death certificates.

Here in Minnesota, where the State has no legal definition of "Negro" several children have birth certificates without "race" designated. Their parents refused to permit their designation as other than American. The U. S. Bureau of Vital Statistics, which will soon discuss the questions which are to be included on its 1966 forms, is credited with requiring racial designation on its reports. However, when inquiry was made by the State of Minnesota, this State was informed that the state has sole jurisdiction over what is recorded on birth certificates. Rather than have a court case, the Minnesota Health Department permitted the certificates in question to be issued with "race" not designated.

If you will pardon a personal reference; in each instance when I filled the selective service forms I wrote "American" in the space where "race" was requested. When the reviewer looked over my academic training (in my absence) he knew that I was "white" and he rewrote my identity as such.

The following is quoted from the U. S. Census for 1940:

> "Three major race classifications are distinguished in the tabulations, namely, white, Negro, and 'other races'. Persons of Mexican birth or ancestry who were not definitely Indian or of other non-white race were returned as white in 1930. ... In 1930 publications, the figures for the white population for 1930 excluded Mexicans, but the 1930 data for whites published in this volume (1940) have been revised to include Mexicans."

The 1920 Census carries the following: "The term 'white' as used in the Census reports refers to persons understood to be pure blooded whites." The identity of the Negro is "according to his racial status in the community in which he lives." In reality "race" in the Census is determined by the enumerator who does not ask the identity of the family but, instead, uses his judgment.

As a professional research worker for over a quarter of a century, and with possibly more field studies under my direction than any other writer dealing with the problems of the "Negro", I fail, at this date, to see the contribution that any public agency in these United States makes to the general welfare by gathering "racial" statistics which set apart one group (and in reality only one group) from the body politic. This nation is constituted of more than a hundred national and ethnic groups. Their inter-mixture is older than the nation itself.

I leave the discussion of moral and religious obligations to others more qualified than I am. Surely you realize the implications here.

Eventually, if not already, the time will come when every citizen will be considered an American and be so designated if for no other reason than that external pressures will force upon us this gesture of solidarity.

SELECTIVE RACIAL STATISTICS

Henry Lee Moon, National Association for the Advancement of Colored People

The National Association for the Advancement of Colored People has not adopted a definitive, overall position on the collection, use and publication of racial data. Within the organization, as within the Negro community generally, there are varying attitudes on the advisability of compiling and publishing such information. The views expressed here conform with the practices of the Association in the absence of a formally declared policy.

The NAACP was founded in 1909 to combat racial segregation, a major obstacle to the progress of the Negro in the United States. One form of segregation is the racial breakdown of statistical information. Nevertheless, the Association from the beginning recognized the need for certain of this information as an instrument in the struggle against segregation. Dr. W.E.B. Dubois, the Association's first director of research, early initiated the collection, publication and interpretation of racial data. On the other hand, the segregationists have effectively employed certain data to retard progress toward the NAACP goal. That goal is the attainment of a society in which considerations of race and color are insignificant, in which the white majority ceases to think of Negroes as a special, different and inferior people, and in which Negroes are no longer compelled to regard themselves as a group apart. In such an ideal society, there would be no thought of classifying data racially.

In his editorial column announcing this panel (The American Statistician, April, 1962), Dr. Philip M. Hauser reviews the pros and cons of the issue and reaches the conclusion "that it is an act of professional integrity and consistency for the statistician to oppose any effort to suppress any form of knowledge about man or the social order."

Both the case for and that against were explored at some length at the Institute on the Collection and Use of Data Based on Race, Religion or National Origin (Philadelphia, Nov. 18, 1959). At that Institute, as in his column, Dr. Hauser set forth three bases in support of the compilation of such data and two against.

To Dr. Hauser's two grounds against, a third may be added: irrelevancy. There is no valid reason for including race on a driver's license, or on any other license issued by a governmental agency. This item has been removed from drivers' licenses issued by the State of New York.

Dr. Hauser cites "the fear that such information may be used to deal with the person or the group in a categoric way, and thus evoke discriminatory practices." This fear is well founded, and is widely shared by members of minority groups. It is the heart of the NAACP opposition to the collection of certain kinds of racial data, even though it is realized that racial statistics do not cause the discrimination; rather they are the product of longstanding discriminatory attitudes and practices. Elimination of such data would not make the Ku Klux Klan or the White Citizens' Councils hate Negroes less. But it would deprive them of official information with which to document their biases.

Take the FBI's annual Uniform Crime Report for the United States which compiles the number of arrests by race. There may be no sound basis on which to challenge the authenticity of the FBI data. Nevertheless, these reports are objectionable on two grounds: first, that crime is an act of the invividual, save when committed by mobs, as lynchings in the United States, or by the state, as genocide in Nazi Germany; and secondly, that the figures are distorted and misleading in that they reveal nothing of the vital socio-economic and psychological factors which restrict the life of every Negro in this country.

Negroes are more easily and frequently arrested and more often convicted than white persons. In part, this is because of their low economic status. The police prey on the poor. Moreover, the poor do not have the means to secure competent legal counsel or to exert political or other influence to get them out of difficulties with the law. In addition, the Negro is all too often the victim of bias in law enforcement and in the administration of justice. Poverty and prejudice thus tend to inflate the figures for Negro arrests, convictions and prison population.

Of a total of 75,385 gambling arrests in 1961, the Uniform Crime Report lists 53,155 Negro arrests as compared with 17,630 white. Can anyone seriously believe that this represents a true picture of gambling in the United States today? That an underprivileged 10 per cent of the population accounts for more than 73 per cent of the country's gamblers? The figures merely reflect the Negro's high vulnerability to arrest.

Equally misleading are the statistics on illegitimate births so widely cited as evidence of low moral standards among Negroes. They show a rate among Negroes ten times greater than that among whites. What they do not show is the relative frequency of abortion or of the use of contraceptives or of the misreporting of legitimacy status. To pass judgment on the sex morality of Negroes on the basis of these figures without reference to comparable data on contraception, abortion and misreporting is gross distortion.

The Kinsey Institute study, Pregnancy, Birth and Abortion, reports a much lower rate of abortion and use of contraceptives among Negro women than among white. In a recent series, The New York Post estimated that there are 1,500,000 illegal operations performed annually in the United States. What proportion of this number is

performed on Negro women is not revealed. In view of the high costs of such operations, it may safely be assumed that relatively few Negroes can afford them.

According to Vital Statistics of the United States, 1959, Volume I, socio-economic factors "are generally believed to operate in the direction of a proportionately greater understatement of illegitimacy in the white race than in the non-white..."

Federally collected and published data on crime and out-of-wedlock births have been used, with devasting effect, by segregationists in their campaign to discredit the Negro and maintain the status quo in race relations. These figures have been widely circulated not only in the South, but also throughout the country as an argument against compliance with the United States Supreme Court's anti-segregation decrees and for the maintenance of other racial barriers. One national magazine editor cites them as a basis for what he calls "rational, valid reasons" for racial prejudice.

That organized bigots and the hate press should take great glee in publicizing these official figures and use them to confirm their ancient prejudices surprises no one. But that reputable periodicals should publish them as naked "facts," as if they existed in a social vacuum and, on this basis, presume to lecture Negroes on their behavior is shocking. John Fischer's editorial in Harper's Magazine (July, 1962) is a case in point. He uses "Negro crime" and "moral irresponsibility" as justifications for white prejudice and calls upon Negro leadership to eradicate these evils. Time Magazine (April 21, 1958), in a similar article, had the grace to concede that "... even heroic efforts by Negro leadership could only dent the Negro crime problem, because essentially it is a white problem. And it will remain a severe problem until Northern whites, private citizens as well as civic officials, recognize that Negro crime is basically a symptom of a failure in integration..."

The NAACP opposes the compilation and publication of racially classified data on crime and illegitimate births because such information sheds no significant light on the causes, because it serves no useful purpose in curbing these offenses, because it is subject to distortion and misrepresentation, and because it is utilized to thwart the drive toward an egalitarian, pluralistic society.

In rejecting these and similar racially classified statistics, the NAACP does not call for the abolition of racial breakdowns in all fields. The Association, in fact, agrees with Dr. Hauser that in certain areas there is a "need for such information for use in 'social engineering' ... as a basis for effectively and efficiently dealing with a wide range of problems ranging from marketing to public policy."

Since the NAACP is engaged in "social engineering" it must have certain racial data which can most efficiently be compiled by a governmental agency. A host of facts revealing the relative socio-economic status of the Negro is needed: how many Negroes there are in the country and in each of the states and subdivisions thereof; their age and sex distribution; the range of their income and what occupations they follow; which industries and companies employ them, in what positions and how many; how they spend their money; what kind of housing is available to them; their educational level and health status; and how many are registered voters. Periodic reports in these areas are needed in order to measure change.

The NAACP is not alone in the need for this information. The Negro press and radio are vitally interested in income and purchasing habits. They need these figures in their solicitation of advertising. Politicians, no less than the NAACP, are concerned about the number and location of Negro voters and what issues influence their ballots.

Employment figures are needed to prove or disprove discrimination and to furnish basic information for FEPC legislation or for court action. Similar information is required in the field of housing where discrimination is widespread. Such documentation aids the fight for enactment of fair housing statutes and for the issuance of an Executive Order banning discrimination in all federally-aided housing.

In the spring of 1962, the New York State Commissioner of Education ordered a racial census of every public school in his jurisdiction. The results revealed the extent of de facto segregation existing in school systems throughout the state and provided an essential factual basis for development of a desegregation program.

To forego the collection and use of such valuable information would be to handicap immeasurably the struggle to rid the nation of racial segregation. The attainment of a color blind society requires an awareness of existing color patterns. The absence of statistical information, racially classified, does not assure the absence of color discrimination. Indeed, it may be used to conceal the color bar behind a facade of impartial color blindness.

The NAACP's unformulated but pragmatic position on the collection and use of racial data is simple. Where such data serve no useful purpose, stop compiling them; where they may be used for constructive ends, continue to collect and publish them. In a word, the Association is for the selective collection, use and publication of racially classified data.

RACIAL CLASSIFICATION IN THE CENSUS

Henry D. Sheldon
Bureau of the Census

The racial classification used in the census has, in the main, evolved in response to matters of public policy rather than in response to the deliberations of advisory committees of racial experts. The "three-fifths" compromise which took official form in Article I, Section 2 of the Constitution focused attention on the Negro population, and separate figures for Negroes were obtained in the first census and at each census thereafter. This classification, presumably, was considered satisfactory until 1860.

In 1860, a "Chinese" category was added to the classification, reflecting in all probability, the public interest in the Chinese immigration of the 1850's and the beginning of anti-Chinese agitation which led to the exclusion act of 1882. Also in the publications of the 1860 Census a category of "Indian" appeared for the first time. It is clear that the Indians had been a problem to our forefathers since earliest times, but it is possible that the language of the Constitution, "excluding Indians not taxed", defined Indians as an irrelevant element in the population to be enumerated. The count was confined to Indians existing as a part of the economy (i.e., taxed), and it was not until 1890 that Indians living on reservations or in "tribal relations" were included in the population of the United States. In 1870, the category "Japanese" was added to the classification--certainly not in response to heavy immigration from Japan--47 Japanese were counted in this census. It seems likely that the category Japanese was included merely as a logical extension of the Chinese category in the previous census.

No further expansion of the racial classification was made until the census of 1910, when in addition to White, Negro, Indian, Chinese, and Japanese, Filipino and a residual "other races" category were added. In that year, 3,015 persons were counted in this residual category, 2,545 Hindus, 462 Koreans, and 8 Maoris. This tendency toward the proliferation of what loosely might be described as Oriental categories continued through the censuses of 1920 and 1930.

In 1930, a further category "Mexican" was added to the racial classification, presumably in response to a general concern relating to the heavy immigration from Mexico during the 1920's. In the years immediately following, the Mexican Government took some exception to this classification, immigration from Mexico declined to a low level, and "Mexican" as a racial category disappeared from the census classification.

It would appear then that, in general, the census racial classification developed in response to issues arising largely in the field of immigration policy, in terms which were comprehensible to, on the one hand, enumerators and, on the other, the general public, including members of Congress. Whether or not it is a racial classification is, of course, an open question. Implicitly, it is, in the sense that the categories are treated as immutable characteristics. That is, persons of Japanese descent are always, according to the classification, Japanese, no matter how many generations in this country; whereas persons born in Germany are identified, and those whose parents were born in Germany are identified, but subsequent generations of German descent are merged in the general population. It is clearly not "scientific" if by scientific we mean conformity to some standard text book classification; and, if such a procedure was desirable, one would be hard put to decide which classification.

Since 1930, public interest involving questions of race has, in the main, shifted from immigration to the field of race relations. There have been Supreme Court decisions, anti-discrimination legislation at both the State and Federal level, and substantial strides in the direction of integration. One phase of this trend has, of course, been the elimination of questions on race on application forms and other records, the discontinuance of the color classification in the compilation of school statistics in many areas, and the elimination of certain types of references to color in the press and in commercial entertainment.

As a part of this general trend, it has been suggested that the item on race be removed from birth and death certificates, and this suggestion has been given serious consideration in a few States. This item was actually removed from these records in the State of New Jersey but has recently been restored. There are, however, some half dozen States in which it has been removed from the face-sheet of the records and buried in the confidential medical part of the record.

There have, likewise, been suggestions that a question on race is inappropriate to the decennial census. This was a matter of some concern to the staff of the Bureau prior to the 1960 Census, since the mailing of the Advance Census Report to every household in the country openly placed the question on race before the public. In previous censuses the enumerator had been instructed to classify the respondent by race without asking the question, except when he was uncertain about the proper classification. With the exception of American Civil Liberties Union, which took occasion to denounce the question on race about a week after the 1960 Census was in the field, there was little opposition to the question. There were no pockets of mass refusals to answer the question, and examination of the Advance Census Reports (filled out by the respondent) showed little, if any, adverse marginal comment. There were several dozen letters of complaint about the question, but the number of such letters was no greater than

those complaining about the income question or housing questions on plumbing as invasions of privacy. In short, the evidence seems to suggest a profound indifference on the part of the general public to the presence of a question on race in the census schedule.

The Bureau's experience with a question on religious preference asked in the Current Population Survey of March 1957 was essentially the same. There were few refusals or complaints from respondents about the question, and, in general, the respondents were not confused by the many elements of potential ambiguity our learned advisory group found in the question.

It would appear then that the opposition to the question on race is more professional than grass roots and that, in the present climate of opinion, a question on race of the traditional census variety is feasible even on the basis of self-enumeration.

XII

EDUCATIONAL STATISTICS

Chairman, Virgil R. Walker, U. S. Office of Education

Page

Some Comparisons of Office of Education and Census Bureau Statistics on Education - Charles B. Nam, U. S. Bureau of the Census..... 258

Some Comparative Effects of Personalized and Nonpersonalized Procedures in Surveys Mailed to Teachers - Glen Robinson and Simeon P. Taylor, III, National Education Association..... 270

A New Method of Probability Sampling in the Field of Education (Abstract) - Philip Desind, U. S. Office of Education..... 278

SOME COMPARISONS OF OFFICE OF EDUCATION AND CENSUS BUREAU STATISTICS ON EDUCATION

Charles B. Nam, Bureau of the Census

Sound educational planning and proper evaluation of the output of the school systems rest, to a considerable extent, on the adequacy of factual knowledge about the educational enterprise. Through statistics we are able to gauge such facts as the extent to which formal educational programs serve the population, the educational status and relative progress of children and young adults in school, and the number of persons expected to pass through the school systems in future years. It is especially important, therefore, that the data which are used in this planning and evaluation are reliable, internally consistent, and relevant to the problem at hand.

The purpose of this paper is to examine critically the correspondence of educational data gathered by the two largest collectors of such information in the United States, namely the Office of Education and the Bureau of the Census. More specifically, comparisons will be made between selected series of data on school enrollment and educational attainment from the Biennial Survey and annual fall surveys of the Office of Education and the decennial census and Current Population Survey of the Census Bureau. [1] How consistent are these series of data? What accounts for differences that do exist? What implications do the findings based on these comparisons have for the application of education data?

Before undertaking these comparisons, it might be helpful to review briefly the major differences which might account for variations in the data, such as in data-collecting techniques, program objectives, and definitions employed by the data collectors.

1. Quite different bases for collecting the data are used by the two agencies. The Office of Education statistics are generally collected by mail questionnaires sent to the officers of State school systems and of institutions of higher education. The Current Population Survey data are collected by household interviews in which the respondent (most often a housewife) provides information about all of the residents. The 1960 Census data were collected partly by self-enumeration and partly by household interview.

2. In both agencies, the statistics on students (which are of most concern in this paper) are gathered in conjunction with statistics on other items. In the Office of Education studies, although the whole survey may be related to educational information, data on teachers, classrooms, and school finances are often collected in the same questionnaire as are the student data. In the Census Bureau censuses and surveys, the education information is collected at the same time as other information about the population, such as marital status, labor force status, income, and so forth. The Office of Education data are used to describe the status of the school systems, its staff, and pupils; the census data are intended to describe the educational status of the population, including those out of school as well as those in school.

3. The adequacy of the data-collection process is difficult to assess. For each data series, there are numerous possible errors arising from this source. The States which compile information for the Office of Education must summarize the reports from local school districts. The local data-collection forms vary considerably in content and language, and some estimation is involved where data are missing or where a local survey has not been taken. Furthermore, although the Office of Education provides the States with standard definitions, deviations from prescribed definitions and instructions are known to exist. In the Census Bureau censuses and surveys, the respondent may sometimes lack complete knowledge about other household members or may sometimes have erroneous information. Also, while interviewers are provided with standard definitions, it is often difficult to transmit this knowledge to the respondent or to have the many respondents interpret the definitions in the same way. Moreover, in the case of the Current Population Survey, which is based on a small national sample, and, to a lesser extent, in the decennial census, which used a 25-percent sample for education data, the statistics are subject to errors of sampling variability.

4. The questions or items used to collect the information in the different sources vary somewhat in definitions employed and time references used. For example, Census Bureau data on college enrollment are intended to refer to all students in college programs which may lead to a degree, whereas Office of Education statistics on college enrollment exclude some types of junior or community college students, such as those in terminal-occupation programs. Also, decennial census data are collected in April and refer to enrollment since February 1 of the year (generally the spring semester), whereas the CPS and Office of Education survey data on college enrollment are collected in the fall of the year. These and other such definitional differences are discussed further in later parts of the paper.

5. In all of the sources, variations may be introduced because of errors in processing or editing the data. Both agencies review the assembled information for reasonableness and internal consistency. Using different techniques, both agencies supply missing information, either by referring to an additional source or by some estimation procedure.

6. There are, of course, other reasons for differences in the data collected, such as different population or geographic coverage and different ways of presenting the data in

published form. These several differences must be considered carefully when comparisons of different sets of data are made.

For the most part, the Office of Education and Bureau of the Census collect different kinds of education data; however, there are some types of information which both agencies collect and which may be compared. I have chosen, for illustrative purposes, four kinds of comparisons. First, public elementary and secondary enrollment by State, according to the 1960 Census, the 1959 O.E. fall survey, and the 1959-60 O.E. Biennial Survey, are examined. Second, public school enrollments by grade, according to the census and the Biennial Survey, are compared. Third, comparisons of public and private college enrollment, based on the 1960 Census, the Current Population Survey, and the O.E. fall survey of higher education, are made. Finally, some attention is paid to the correspondence of census data on educational attainment and O.E. data on high school graduates and college degrees.

As seen in table 1, public school enrollment (that is, in kindergarten through the twelfth grade) was 35.3 million in the 1960 Census and 35.2 million in the 1959 O.E. fall survey. One might have expected the census figure to be slightly lower because the data refer to the spring semester and some slight attrition at these grades takes place between the fall and spring terms. However, some shifting from parochial or other private schools to public schools probably also takes place during that time, and this shifting would compensate for the attrition effect. At any rate, the figures are quite close, with the census total being only slightly over 100 percent of the O.E. figure. Looking at the comparison by States, one sees some greater discrepancies. Nevertheless, in only 13 States is the deviation as great as 3 percent and in only one State (Alaska) is the deviation extreme. 2/ Most of these differences, moreover, can probably be accounted for, in great part, to varying definitions of residence or to transfers or residential mobility between States between the fall and spring. Some information to test this hypothesis will be available in later census reports.

The differences between the O.E. Biennial Survey figures and either of the other two sources are greater than the differences between those two sources. Here, obviously, the main factor is the nature of the Biennial Survey data, which are a cumulative count of all children enrolled in the State during the school year. Because of transfers, some students may enroll in the school systems of two or more States during the year. Even so, the Biennial data show a marked tendency to correspond, in level, to the census and O.E. fall figures.

The discrepancies between the 1960 Census statistics and the O.E. Biennial Survey data are of about the same order of magnitude, as seen in table 2, when a division is made between the elementary and secondary level (that is, between kindergarten through the eighth grade and the ninth through the twelfth grades). At the national level, the differences is a bit greater at the elementary then at the secondary level. Figures from the Census Bureau's Current Population Survey for the fall of 1959 fit in generally well with the decennial census and O.E. Biennial figures. At the elementary level, the CPS figure falls below the other two series (possibly, in part, because four-year-olds, some of whom are in kindergarten in the fall, are not included in the age coverage for enrollment in the CPS). At the secondary level, the CPS figure falls exactly between the census and O.E.figures.

Shifting to a comparison of grade distributions in the census and the Biennial Survey, as shown in table 3, we find exceptionally good correspondence except at the first and twelfth grades. The reasons for the differences at these grades are somewhat difficult to determine; however, there are some clues. The Biennial figure, which is higher at the first grade, is a cumulative count, and it is possible that duplication is greatest at the first grade where the number of grade repeaters is high. On the other hand, because of the way the decennial census data on grade in which enrolled are collected and tabulated (that is, by combining responses to a question on school enrollment with responses to a question on highest grade attended, where net overreporting of the latter is generally found), it may be that errors of misreporting in the census are greatest at the first and twelfth grades (at the latter grade because of some erroneous inclusion of older persons in adult education courses). 3/ Some support for the latter hypothesis was discovered by extending the comparison to include the grade distribution from the fall 1959 Current Population Survey in which grade enrollment is obtained by a more direct question than in the census. The CPS figures, which are not shown in the tables, are clearly much closer to the Biennial Survey than to the decennial census figures, including comparisons at the first and twelfth grades.

Comparisons of college enrollment by State can be made between the 1960 Census and the 1959 O.E. fall survey of higher education. According to table 4, the census figures for the United States as a whole and for most of the States are considerably below those from the O. E. survey. At the national level, the census figure is 13 percent less than the O.E. figure. There is a great deal of variation in the relationship by States. The primary reason for these differences, which incidentally are very similar to those between the 1950 Census and the 1949 O. E. fall survey, probably can be found in the time of the year to which the data refer. Considerable attrition in college enrollment takes place between the fall and spring semesters. A special inquiry concerning fall and spring enrollment, sent to the largest universities in six States where

the census figure fell appreciably below the O.E. figure, showed that the attrition rate was very close to the difference between the two sets of data. 4/ Further confirmation of this explanation is found by comparing the national figure on college enrollment obtained from the fall 1959 Current Population Survey with the O.E. fall estimate. The CPS figure, which was 3,340,000, is remarkably close to the 3,377,273 number resulting from the O.E. fall survey, especially in view of the slight definitional differences and sampling variability in the CPS.

The college enrollment statistics may be compared further according to whether enrollment was in a public or private college, as in table 5. This information was collected in a decennial census for the first time in 1960 but has been collected over a period of years in the Current Population Surveys and in the O.E. fall surveys. The percentage of college students in public colleges consistently runs somewhat higher in the CPS than in the O. E. survey. Because of this difference, a check was made on the CPS reports in 1958. Respondents were asked in the usual way whether a college enrollee was in a public or private college and, at another point in the interview, were also asked for the name of the college being attended. Classification of the college names was made by using the Office of Education college directory and comparaisons were made on a case-by-case basis with the CPS report. The results are shown in table 6. If, as seems proper, one accepts the O. E. classification, the CPS reports which were in error were largely those in which a private college was reported as public. Since many private colleges have place names (such as the University of Dayton), this particular type of error is understandable. Some further instructions to CPS interviewers in 1959 on this matter did not result in improved reporting, however, as can be seen in table 5; but the 1960 decennial census figures matched exactly, percentagewise, with the O. E. figures. The greater census accuracy in reporting this item probably can be explained because college students were counted in the census where they attended school and they tended to report for themselves, whereas in the CPS unmarried college students are reported by their parents whose knowledge of the type of control of the college may have been incomplete.

As a final illustration, I shift from an examination of enrollment statistics to an examination of educational attainment statistics. There are few direct comparisons between Census Bureau and Office of Education data of this type which can be made but it is possible to make some indirect comparisons. For instance, according to the March 1959 Current Population Survey, about 6.9 million persons 25 to 29 years old were high school graduates (that its, had completed four years of high school or some college); if one sums the number of high school graduates each year reported by the Office of Education for a five-year period corresponding to the time at which this age group should have completed high school, about 6.0 million persons are recorded as high school graduates. The most obvious reason for this difference is the probable net overreporting of persons completing high school, according to the CPS; however, it is difficult to match cohorts in the two sets of data, and some persons receive high school equivalent certificates at a late age. It is significant to note that in a special study conducted by the Census Bureau, in which information was gathered about youths, reported as high school seniors in the October 1959 CPS, as to their high school status in the fall of 1960, an estimated 1,803,000 were reported as high school graduates; and this number agreed exactly with the independent estimate made by the Office of Education based on their survey.

When the number of college graduates (that is, the persons who had completed four or more years of college, according to the CPS) for the same 25-to-29 year age group is compared with the number of college degrees (bachelors and first professional level, according to O.E.) for the appropriate time period, the CPS figure is found to be slightly lower than the O.E. figure (about 1.2 vs. 1.4 million). Quite likely, some net overreporting of college graduates in the CPS was more than compensated for by the fact that the O.E. data included many persons (maybe up to 10 percent of the total with bachelor or first professional degrees) who received a professional degree after having received a bachelor's degree in the same five-year period. Also, here again, the difficulty of exactly matching cohorts is vital.

To sum up, although the illustrations shown here cover only some of the comparisons which could be made, and although probably not all of the differences found can be rationalized, it seems reasonable to conclude that, allowing for fundamental variations in definitions, coverage, time references, and the like, there is remarkable correspondence between Office of Education and Census Bureau statistics on education. However, although this finding can be, and is, comforting to those of us who produce the data, the main point of the paper goes beyond this discovery.

The fact is that there <u>are</u> differences in the several sets of data and these differences are there largely because the different data producers rightly structure the collection of data to meet the needs of their major consumers. Therefore, unless the nature of the data and the reasons for differences in the sets of statistics are considered by the data users, serious errors in application and interpretation might be the result. This argues, I think, for two recommendations: (1) Consumers of these education statistics should not take too lightly the differences in the data and should not use the sets of data interchangeably without qualification. This requires, I believe, some type of educational program. (2) Perhaps, to some degree, the sort of careless misuse or misinterpretation of education data which is often

heard in some speeches and found in some publications is attributable to a lack of emphasis by the data producers on the nature and quality of the data. I am not sure what the best approach is here, but it is incumbent upon both producers and consumers to see that statistics which are collected and published are properly presented, correctly analyzed and interpreted, and judiciously applied.

FOOTNOTES

1/ The Census Bureau also gathers data on public school organization and finance through its Government Division's program, but no consideration is being given to that source in this paper. Likewise, no comparisons are being made in the area of teacher statistics or of other education items for which both agencies collect data.

2/ The Office of Education does not count some schools in Alaska as regular public schools which would be so regarded in the census.

3/ Some recent analysis of census and O.E. grade distributions in 1940, 1950, and 1960 shows greater convergence in the two series at each later date.

4/ This suggests that the extent of underenumeration of college students in the census may have been small. A proper evaluation of this point, however, must await data from the 1960 Census evaluation program, which includes a specific check on the coverage of college students.

Table 1.--PUBLIC SCHOOL ENROLLMENT IN KINDERGARTEN THROUGH THE TWELFTH GRADE, IN THE UNITED STATES, BY STATE: 1960 CENSUS, 1959 OFFICE OF EDUCATION FALL SURVEY: AND 1959-60 OFFICE OF EDUCATION BIENNIAL SURVEY

Area	Public school enrollment, grades K to 12			Census as percent of--	
	1960 Census 1/	1959 OE Fall Survey 2/	1959-60 OE Biennial Survey 3/	1959 OE Fall Survey	1959-60 OE Biennial Survey
UNITED STATES.....	35,282,104	35,182,343	36,142,790	100.28	97.62
Alabama.....	781,605	775,931	787,269	100.72	99.28
Alaska.....	47,739	39,449	44,450	121.01	107.40
Arizona.....	292,685	280,369	302,182	104.39	96.86
Arkansas.....	419,085	422,060	424,206	99.30	98.79
California.....	3,256,081	3,190,745	3,199,455	102.05	101.77
Colorado.....	378,990	371,608	393,679	101.99	96.27
Connecticut.....	468,607	460,311	476,749	101.80	98.29
Delaware.....	79,095	77,761	80,874	101.72	97.80
Dist. of Columbia.....	114,910	116,587	122,446	98.56	93.85
Florida.....	967,867	919,491	993,496	105.26	97.42
Georgia.....	939,641	930,624	949,099	100.97	99.00
Hawaii.....	144,082	140,491	139,429	102.56	103.34
Idaho.....	163,222	155,890	162,839	104.70	100.24
Illinois.....	1,689,995	1,683,442	1,787,869	100.39	94.53
Indiana.....	935,013	940,000	988,872	99.47	94.55
Iowa.....	565,669	567,253	598,100	99.72	94.58
Kansas.....	453,607	453,835	478,446	99.95	94.81
Kentucky.....	614,054	610,992	631,412	100.50	97.25
Louisiana.....	696,418	693,202	693,199	100.46	100.46
Maine.....	197,950	197,035	195,214	100.46	101.40
Maryland.....	578,810	583,015	596,363	99.28	97.06
Massachusetts.....	855,184	831,230	859,865	102.88	99.46
Michigan.....	1,623,903	1,609,893	1,625,247	100.87	99.92
Minnesota.....	672,952	682,306	681,358	98.63	98.77
Mississippi.....	565,051	560,000	566,421	100.90	99.76
Missouri.....	756,364	805,000	820,724	93.96	92.16
Montana.....	143,662	142,850	144,973	100.57	99.10
Nebraska.....	282,957	274,262	282,719	103.17	100.08
Nevada.....	63,053	60,279	66,415	104.60	94.94
New Hampshire.....	103,897	102,745	105,773	101.12	98.23
New Jersey.....	1,033,711	1,015,000	1,050,992	101.84	98.36
New Mexico.....	225,387	211,334	231,004	106.65	97.57
New York.....	2,727,540	2,697,477	2,828,853	101.11	96.42
North Carolina.....	1,072,045	1,082,213	1,105,412	99.06	96.98
North Dakota.....	141,238	130,173	136,766	108.50	103.27
Ohio.....	1,852,762	1,832,479	1,905,995	101.11	97.21
Oklahoma.....	510,751	533,928	533,489	95.66	95.74
Oregon.....	380,938	379,699	388,772	100.33	97.98
Pennsylvania.....	1,894,009	1,948,985	1,953,661	97.18	96.95
Rhode Island.....	129,177	128,406	133,155	100.60	97.01
South Carolina.....	581,364	573,084	610,099	101.44	95.29
South Dakota.....	155,275	147,778	153,596	105.07	101.09
Tennessee.....	780,003	779,928	810,300	100.01	96.26
Texas.....	2,016,349	2,084,540	2,068,158	96.73	97.49
Utah.....	230,260	228,826	235,934	100.63	97.60
Vermont.....	74,647	72,822	72,790	102.51	102.55
Virginia.....	808,668	820,578	841,574	98.55	96.09
Washington.....	616,339	609,024	608,999	101.20	101.21
West Virginia.....	428,658	451,875	460,429	94.86	93.10
Wisconsin.....	692,746	700,000	732,238	98.96	94.61
Wyoming.....	78,089	75,488	81,431	103.45	95.90

Table 1 continued

1/ Refers to enrollment at any time between February 1 and April 1, 1960.

2/ Refers to full-time enrollment "on or about October 1, 1959, or the nearest date thereto when enrollment can be considered stabilized."

3/ These "State enrollment data generally represent a cumulative count of the total number of different pupils registered at any time during the school year in each State."

Note: An estimate of public school enrollment in grades K to 12 in the United States, according to the Census Bureau's Current Population Survey in October 1959, was 34,929,000.

Sources:

U.S. Bureau of the Census. U. S. Census of Population: 1960. General Social and Economic Characteristics. Final Reports PC(1)-C, Parts 1-52.

U. S. Office of Education. Fall 1960 Statistics on Enrollment, Teachers, and Schoolhousing in Full-Time Public Elementary and Secondary Day Schools. Circular No. 634, 1960. (Includes revised 1959 data.)

U. S. Office of Education. Preliminary Statistics of State School Systems, 1959-1960. Circular No. 663, July 1961.

U. S. Bureau of the Census. "School Enrollment: October 1959," Current Population Reports, Series P-20, No. 101, May 22, 1960.

Table 2.--PUBLIC SCHOOL ENROLLMENT IN KINDERGARTEN THROUGH THE EIGHTH GRADE AND NINTH THROUGH THE TWELFTH GRADE IN THE UNITED STATES, BY STATE: 1960 CENSUS AND 1959-60 OFFICE OF EDUCATION BIENNIAL SURVEY

Area	Grades K to 8			Grades 9 to 12		
	1960 Census	1959-60 OE Biennial Survey	Census as percent of OE	1960 Census	1959-60 OE Biennial Survey	Census as percent of OE
UNITED STATES...	26,666,656	27,629,881	96.51	8,615,448	8,512,909	101.20
Alabama.....	592,859	608,563	97.42	188,746	178,706	105.62
Alaska.....	38,378	36,820	104.23	9,361	7,630	122.69
Arizona.....	225,673	235,717	95.74	67,012	66,465	100.82
Arkansas.....	311,607	323,789	96.24	107,478	100,417	107.03
California.....	2,452,623	2,473,392	99.16	803,458	726,063	110.66
Colorado.....	287,656	308,870	93.13	91,334	84,809	107.69
Connecticut.....	357,045	365,947	97.57	111,562	110,302	100.69
Delaware.....	60,132	61,655	97.53	18,963	19,219	98.67
Dist. of Columbia...	90,615	99,173	91.37	24,295	23,273	104.39
Florida.....	734,690	768,025	95.66	233,177	225,471	103.42
Georgia.....	724,484	748,166	96.83	215,157	200,933	107.08
Hawaii.....	106,034	106,008	100.02	38,048	33,421	113.84
Idaho.....	117,128	120,414	97.27	46,094	42,425	108.65
Illinois.....	1,271,699	1,356,360	93.76	418,296	431,509	96.94
Indiana.....	706,817	752,435	93.94	228,196	236,437	96.51
Iowa.....	429,737	456,986	94.04	135,932	141,114	96.33
Kansas.....	340,862	365,565	93.24	112,745	112,881	99.88
Kentucky.....	480,803	496,063	96.92	133,251	135,349	98.45
Louisiana.....	536,303	542,379	98.88	160,115	150,820	106.16
Maine.....	149,785	151,856	98.64	48,165	43,358	111.09
Maryland.....	440,591	457,975	96.20	138,219	138,388	99.88
Massachusetts.....	638,434	643,044	99.28	216,750	216,821	99.97
Michigan.....	1,246,974	1,253,466	99.48	376,929	371,781	101.38
Minnesota.....	487,858	503,845	96.83	185,094	177,513	104.27
Mississippi.....	442,486	451,921	97.91	122,565	114,500	107.04
Missouri.....	567,110	624,744	90.77	189,254	195,980	96.57
Montana.....	105,739	109,164	96.86	37,923	35,809	105.90
Nebraska.....	209,304	215,579	97.09	73,653	67,140	109.70
Nevada.....	46,922	52,242	89.82	16,131	14,173	113.82
New Hampshire.....	78,919	81,064	97.35	24,978	24,709	101.09
New Jersey.....	773,954	796,462	97.17	259,757	254,530	102.05
New Mexico.....	172,532	179,806	95.95	52,855	51,198	103.24
New York.....	2,020,328	2,126,130	95.02	707,212	702,723	100.64
North Carolina.....	821,698	844,346	97.32	250,347	261,066	95.89
North Dakota.....	100,870	101,166	99.71	40,368	35,600	113.39
Ohio.....	1,423,013	1,469,707	96.82	429,749	436,288	98.50
Oklahoma.....	374,604	395,401	94.74	136,147	138,088	98.59
Oregon.....	279,237	288,638	96.74	101,701	100,134	101.56
Pennsylvania.....	1,395,058	1,440,474	96.85	498,951	513,187	97.23
Rhode Island.....	96,955	100,814	96.17	32,222	32,341	99.63
South Carolina.....	456,427	482,442	94.61	124,937	127,657	97.87
South Dakota.....	113,939	116,436	97.86	41,336	37,160	111.24
Tennessee.....	602,984	628,562	95.93	177,019	181,738	97.40
Texas.....	1,559,146	1,611,765	96.74	457,203	456,393	100.18
Utah.....	172,216	178,064	96.72	58,044	57,870	100.30
Vermont.....	57,110	56,725	100.68	17,537	16,065	109.16
Virginia.....	625,960	663,363	94.36	182,708	178,211	102.52
Washington.....	456,244	453,489	100.61	160,095	155,510	102.95
West Virginia.....	327,743	351,591	93.22	100,915	108,838	92.72
Wisconsin.....	499,159	510,988	97.69	193,587	221,250	87.50
Wyoming.....	58,212	62,285	93.46	19,877	19,146	103.82

Table 2 continued

Cf. footnotes on table 1.

Note: According to the Census Bureau's Current Population Survey in October 1959, for the United States, there were an estimated 26,358,000 enrolled in grades K to 8 and 8,571,000 enrolled in grades 9 to 12.

Cf. source notes on table 1.

Table 3.--PUBLIC SCHOOL ENROLLMENT, BY GRADE, IN THE UNITED STATES: 1960 CENSUS AND 1959-60 OFFICE OF EDUCATION BIENNIAL SURVEY

Grade	Numbers		Percent distribution		Census as percent of OE Survey
	1960 Census 1/	1959-60 OE Biennial Survey 2/	1960 Census 1/	1959-60 OE Biennial Survey 2/	
Total, K to 12...	35,282,091	36,146,846	100.0	100.0	97.60
Kindergarten.....	1,846,374	1,923,384	5.2	5.3	95.99
First.....	3,479,028	3,739,471	9.9	10.3	93.03
Second.....	3,345,237	3,445,997	9.5	9.5	97.07
Third.....	3,235,740	3,311,361	9.2	9.2	97.71
Fourth.....	3,035,552	3,145,630	8.6	8.7	96.50
Fifth.....	3,013,021	3,116,819	8.5	8.6	96.66
Sixth.....	2,995,919	3,070,084	8.5	8.5	97.58
Seventh.....	3,046,627	3,170,737	8.6	8.8	96.08
Eighth.....	2,669,149	2,706,398	7.6	7.5	98.62
Ninth.....	2,357,127	2,415,806	6.7	6.7	97.57
Tenth.....	2,215,375	2,264,299	6.3	6.3	97.83
Eleventh.....	2,055,546	2,068,567	5.8	5.7	99.37
Twelfth 1/.....	1,987,396	1,768,293	5.6	4.9	112.39

Cf. footnoes on table 1.

1/ Includes postgraduate high school students.

Table 4.--COLLEGE ENROLLMENT, FOR THE UNITED STATES BY STATE: 1960 CENSUS AND 1959 OFFICE OF EDUCATION FALL SURVEY

Area	College enrollment		Census as percent of Office of Education Survey
	1960 Census 1/	1959 Office of Education Fall Survey 2/	
UNITED STATES.....	2,935,446	3,377,273	86.92
Alabama.....	39,638	46,138	85.91
Alaska.....	1,849	2,767	66.82
Arizona.....	23,394	32,785	71.36
Arkansas.....	21,937	23,529	93.23
California.....	325,918	411,029	79.29
Colorado.....	40,710	45,832	88.82
Connecticut.....	41,914	47,015	89.15
Delaware.....	4,990	6,434	77.56
Dist. of Columbia.....	23,553	44,583	52.83
Florida.....	55,167	64,607	85.39
Georgia.....	47,430	47,564	99.72
Hawaii.....	8,636	9,645	89.54
Idaho.....	11,530	11,082	104.04
Illinois.....	156,883	182,721	85.86
Indiana.....	78,163	89,322	87.51
Iowa.....	48,123	54,063	89.01
Kansas.....	42,262	50,102	84.35
Kentucky.....	38,034	43,652	87.13
Louisiana.....	47,471	53,996	87.92
Maine.....	11,811	11,561	102.16
Maryland.....	48,029	54,095	88.79
Massachusetts.....	118,733	127,779	92.92
Michigan.....	132,157	152,294	86.78
Minnesota.....	65,499	71,777	91.25
Mississippi.....	31,145	33,102	94.09
Missouri.....	66,179	79,360	83.39
Montana.....	11,618	11,923	97.44
Nebraska.....	25,155	31,298	80.37
Nevada.....	3,248	3,708	87.59
New Hampshire.....	11,375	11,556	98.43
New Jersey.....	82,032	82,039	99.99
New Mexico.....	13,660	16,575	82.41
New York.....	313,147	342,208	91.51
North Carolina.....	61,219	66,410	92.18
North Dakota.....	12,656	13,242	95.57
Ohio.....	137,368	164,749	83.38
Oklahoma.....	45,942	55,763	82.39
Oregon.....	33,859	40,421	83.77
Pennsylvania.....	159,777	179,084	89.22
Rhode Island.....	15,196	17,473	86.97
South Carolina.....	28,709	31,232	91.92
South Dakota.....	12,783	13,989	91.38
Tennessee.....	50,474	59,886	84.28
Texas.....	151,412	179,322	84.44
Utah.....	30,023	30,378	98.83
Vermont.....	8,985	8,325	107.93
Virginia.....	50,406	53,325	94.69
Washington.....	54,867	62,350	88.00
West Virginia.....	22,523	28,255	79.71
Wisconsin.....	63,302	70,478	89.82
Wyoming.....	4,555	6,540	69.65

Table 4 continued

1/ Refers to enrollment at any time between February 1 and April 1, 1960 in college programs which may lead to a degree.

2/ Refers to enrollment of degree-credit students "as of the earliest date on which the enrollment becomes stabilized; normally this is within 2 weeks after the opening of the term." Excludes students in technical institutes, terminal-occupational programs, or organized occupational curriculum, even though they may be degree-credit students.

Note: Total college enrollment in the United States, excluding Alaska and Hawaii, according to the Census Bureau's Current Population Survey in October 1959, was 3,340,000.

Table 5.--PUBLIC AND PRIVATE COLLEGE ENROLLMENT IN THE UNITED STATES: 1960 CENSUS, 1959 CURRENT POPULATION SURVEY, AND 1959-60 OFFICE OF EDUCATION BIENNIAL SURVEY

Type of enrollment	1960 Census		1959 CPS		1959-60 OE Survey	
	Number	Percent distribution	Number	Percent distribution	Number	Percent distribution
Total.....	2,935,446	100.0	3,340,000	100.0	3,402,297	100.0
Public.....	1,727,789	58.9	2,120,000	63.5	2,002,868	58.9
Private.....	1,207,657	41.1	1,220,000	36.5	1,399,429	41.1

Table 6.--COMPARISON OF CPS ENTRIES ON PUBLIC OR PRIVATE COLLEGE ENROLLMENT WITH OFFICE OF EDUCATION CLASSIFICATION OF COLLEGES ATTENDED BY CPS RESPONDENTS: UNWEIGHTED CPS SAMPLE, OCTOBER 1958

Category	Number	Percent
Number of college students in sample.....	2,062	-
Name of college not given or incomplete.....	143	-
Total number of cases with college reported.....	1,919	100.0
Number with same classification in CPS and OE.....	1,778	92.7
Number with different classification in CPS and OE....	141	7.3
OE-Public; CPS-Private.....	26	1.4
OE-Private; CPS-Public.....	115	6.0
Net number with different classification.....	89	4.6

SOME COMPARATIVE EFFECTS OF PERSONALIZED AND NONPERSONALIZED PROCEDURES IN SURVEYS MAILED TO TEACHERS

Glen Robinson, National Education Association
Simeon P. Taylor, III, National Education Association

Introduction

Since 1959, the Research Division of the National Education Association has been engaged in a concerted effort to streamline its statistical and data-gathering procedures. This effort, termed the Sampling Project, has brought together the most up-to-date sampling procedures available and is adapting them to teacher, principal, and superintendent populations. The central purpose of the project is to improve the accuracy and speed of data collecting and reporting in nationwide surveys.

A well-known and important factor affecting the accuracy of data collected through mailed surveys is the usually low rate of response and the possible accompanying high nonresponse bias. The initial phase of the Sampling Project focused on this problem. Through careful planning, efficient control of mailing lists, appropriate processing of incoming questionnaires, effective initial contact letters, and effective follow-up procedures, remarkable response rates have been achieved. The mean net response rate (useable questionnaires returned) for seven surveys conducted in 1960 and 1961 was 95.7 percent.1/

In the first phase of the project, the objective was to secure the highest response rates that could possibly be obtained from the populations sampled; therefore, cost was not a factor considered in initial contact and follow-up procedures. The many theories on methods of increasing response rates were examined and the most plausible were incorporated into the survey procedures.

Since it was believed that personalized approaches in initial contact letters and follow-up procedures were essential to achieving high response rates, such approaches were employed to the fullest. For example, survey instruments were accompanied by a personally addressed and typed letter, signed by the Director of the Research Division. This letter explained the nature of the survey and the importance of each teacher's response. This letter contained a copy of the survey instrument and a self-addressed envelope bearing a postage stamp.

The general pattern of follow-up procedures consisted of a personal letter sent via air mail to nonrespondents at the end of two weeks. A telegram follow-up was sent at the end of four weeks, and another at the end of five weeks. Cut-off dates for receipt of retured questionnaires were established at eight weeks following the initial mailing date.

In conducting the surveys, special effort was made to glean as much information as possible about the populations studied and their response patterns. This information has been valuable in analyzing factors associated with nonresponse and the accompanying biases.

The Cost-Quality Problem

The estimate of $214 per 100 persons for typing, postage, and telegraphy used in the highly personalized contact and follow-up procedures of this first phase of the project made the procedures prohibitive for general use in nationwide surveys. The Division, therefore, was faced with the problem of how much the highly personalized approach could be relaxed without causing a significant drop in the response rate. We believed the rate should be kept above 90 percent.

Several questions had to be answered. Among them were:

1. How much of the high response rate, if any, was due to the personalized effect of the telegram?
2. Could the telegrams be replaced by letters and still achieve a response rate above 90 percent within a reasonable time?
3. Were personalized initial contact and follow-up letters essential for a minimum response rate or could nonpersonalized letters be substituted?

To find answers to these questions, an experimental design was developed to test variations in initial contact and follow-up procedures.

Relative Drawing Power

Two hypotheses were formulated for the purpose of testing the relative drawing power of personalized and nonpersonalized contact procedures. These hypotheses were:

1/ For a description of the seven surveys, the procedures used, and estimates of the sampling variability in the studies see: Robinson, Glen, and McCall, Chester H., Jr. "Some Statistical Findings from Nationwide Teacher Polling." Proceedings of the Social Statistics Section, 1961. Washington, D. C.: American Statistical Association, 1961. p. 56-63.

I. Letter follow-up procedures have the power to elicit from samples of teacher populations in mailed surveys a rate of response equal to that elicited by procedures involving telegrams.

II. Personalized initial contact letter and follow-up procedures have the power to elicit a significantly higher rate of response from teacher populations in mailed surveys than do nonpersonalized letters.

The Experimental Design

The NEA Teacher Opinion Polls were selected as the test media. The survey instruments used in these polls consist of a single sheet of paper printed front and back containing approximately 14 opinion questions and 12 status questions. These instruments are mailed to a nationwide probability sample of classroom teachers. Samples have ranged in size from 1,147 to 1,633.

Polls I and II were conducted in the spring of 1960 with almost identical contact and follow-up procedures. These consisted of an initial contact letter and a first follow-up letter personally addressed, typed, signed, and sent air mail, and second and third follow-ups with telegrams. The polls achieved responses of 96.8 percent and 94.9 percent, respectively, within an eight-week period.

Poll III was conducted in the spring of 1961. The follow-up procedure was altered by the substitution of a second personalized letter for the first telegram follow-up. Figure I shows the variations in the response patterns among the three polls. Table 1 shows the cumulative response rates. It should be noted that Poll III achieved a 96.6-percent response within eight weeks. Obviously, the variations among the response rates of Polls I, II, and III could be attributed to chance alone.

When Poll IV was conducted in the spring of 1962, it was decided to depart substantially from the procedures in the previous polls as follows:

1. The nationwide probability sample of 1,464 classroom teachers selected for Poll IV was randomly divided into two groups of 732.[2/]

2. Neither group received telegrams. Both groups received an initial contact letter followed by a postal card reminder at the end of two weeks and four follow-up letters at two-week intervals. The contents of the letters and the postal cards were identical for the two groups. Follow-ups for the two groups were mailed simultaneously.

3. The treatment given the two groups differed in that the contact and follow-up letters sent to Group A were personally addressed to the teacher, typed, and signed by the Director of the Division; whereas the contact letters and the first two follow-up letters sent to Group B were impersonally addressed, duplicated by offset machine, and bore only a facsimile signature. The last two letters sent to Group B were personalized for the reason explained later.

The two purposes for this division were:

1. To compare the results received from Group A with Polls I, II, and III to determine

2/ For description of sampling procedure see footnote 1.

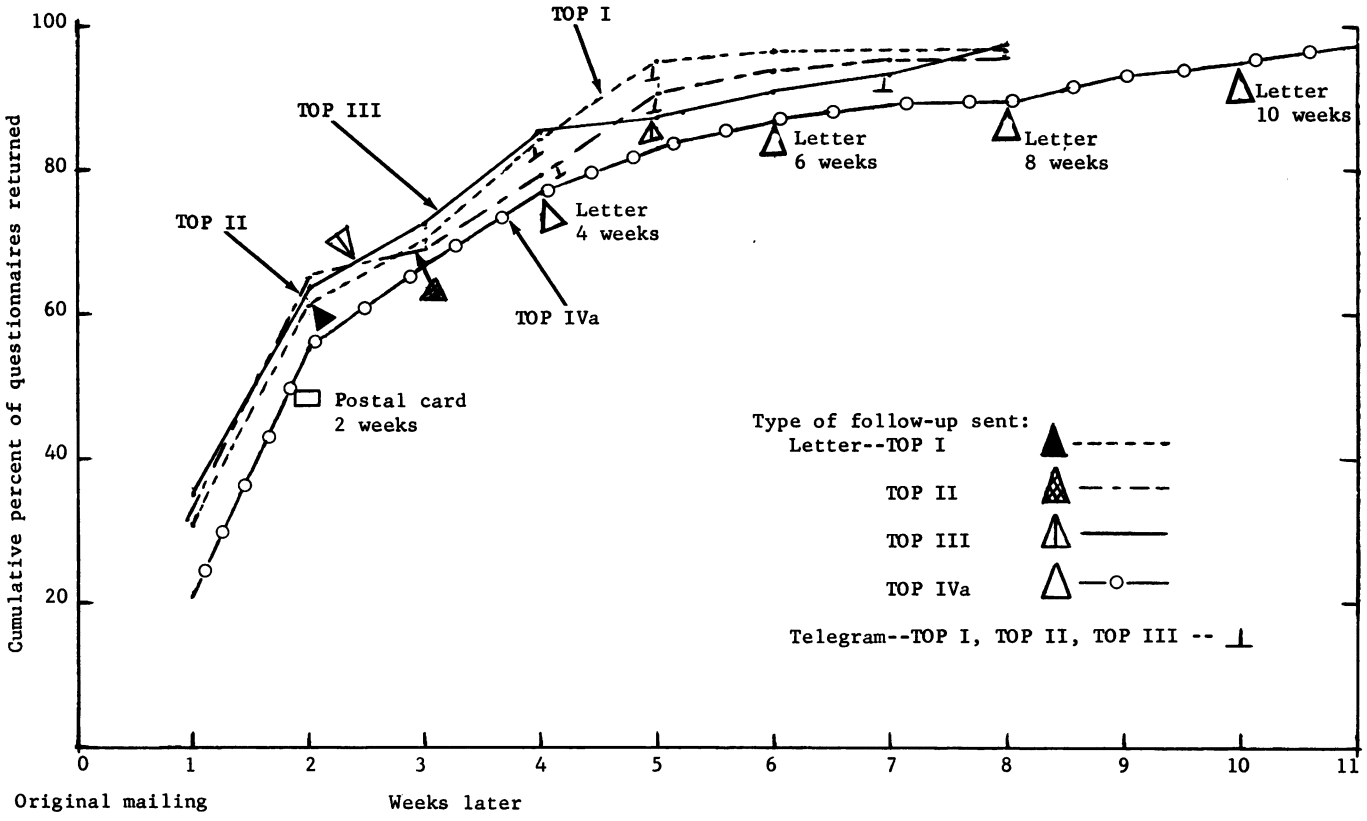
TABLE 1.--CUMULATIVE PERCENT OF RESPONSE[a/] FOR TEACHER OPINION POLLS I, II (1960), AND III (1961), AND FOR TEACHER OPINION POLL IV, EXPERIMENTAL GROUP A (1962)

Weeks after mailing	TOP I n=1,149	TOP II n=1,147	TOP III n=1,633	TOP IV[a/b/] n=732
1	2	3	4	5
1	30.1%	31.5%	35.1%	20.9%
2	61.4	64.8	64.0	55.3
3	71.1	69.4	72.2	66.8
4	84.1	79.7	84.2	76.5
5	95.3	91.2	86.8	82.9
6	96.3	94.0	90.9	86.3
7	96.8	94.9	93.3	88.7
8	96.8	94.9	96.6	89.6
9	...	...	...	93.0
10	...	...	...	94.4
11	...	...	...	96.4

a/ Percents of response are cumulative net responses which include only those persons returning useable questionnaires.

b/ These data are for TOP IV experimental Group A only.

FIGURE I--RESPONSE PATTERNS FOR TEACHER OPINION POLLS I (1960), II (1960), III (1961), AND IVa (1962)



if telegrams could be eliminated completely, and thus serve as a test for Hypothesis I.

2. To have Group B serve as a control for Group A in testing Hypothesis II (personalized vs. nonpersonalized letters).

Experimental Findings

A brief discussion of the findings of these experiments follows:

Telegrams vs. Personalized Letters

Figure I shows the response pattern for Poll IV, and Group A. Table 1 shows the cumulative net response ratios. Comparison of response curves in Figure I indicate how the response to the personalized letter by Group A differs from that obtained in the three previous polls. Utilization of the Kolmogorov-Smirnov test demonstrates that the response distribution for Poll IV, Group A, was significantly different from the other three polls (α=.001). At the end of eight weeks the response rate for Group A was only 89.6 percent compared with a mean response of 96.1 percent for the other three polls. This difference was significant at the .05 level.

But the important thing is that at the end of 11 weeks the response rate was 96.4 percent. This was 1.5 percentage points above the 94.9-percent response in Poll II and only 0.4 percentage points below Poll I, both of which had used two telegram follow-ups. Obviously, variations of this size could easily occur by chance alone.

TABLE 2.--TREATMENT OF GROUPS A AND B IN TEACHER OPINION POLL IV

Action and date	Group A: Experimental, personalized treatment (n = 732)	Group B: Control, nonpersonal treatment (n = 732)
1	2	3
Initial contact February 7	Personalized letter	Nonpersonalized letter
1st follow-up Two weeks	Nonpersonalized postal card reminder	Nonpersonalized postal card reminder
2nd follow-up Four weeks	Personalized letter	Nonpersonalized letter
3rd follow-up Six weeks	Personalized letter	Nonpersonalized letter
4th follow-up Eight weeks	Personalized letter	Personalized letter
5th follow-up Ten weeks	Personalized letter	Personalized letter

TABLE 3.--VARIATIONS IN FIRST EIGHT WEEKS OF FOLLOW-UP

Factor varied	Group A (Experimental)	Group B (Control)
1	2	3
Method of writing	Wording of entire letter written on automatic electric typewriter	Wording of entire letter typed on multilith mats and duplicated on an offset machine
Inside address	Addressed by name and school address to the individual in the sample	No inside address
Salutation	Salutation contained the name of the individual recipient, i.e., Dear Mr. Doe.	All letters bore general salutation of "Dear teacher"
Signature	Personal signature of the director of the NEA Research Division in ink	Facsimile signature of director of the NEA Research Division from multilith mat

Since the final response rate for Poll IV, Group A, was substantially equal to the rates obtained in the previous polls, and since 11 weeks is a reasonable period of time in these studies, Hypothesis I is acceptable. Hence, we conclude that personalized letter follow-up procedures have the power to elicit from teacher populations, within a reasonable time, a response equal to that elicited by procedures involving telegrams.

There is evidence, however, that the use of telegrams in follow-up procedures does reduce the time of response.

Personalized vs. Nonpersonalized Letters

Hypothesis II was tested by a carefully controlled experiment, the general design of which was briefly described previously. A probability sample of 1,464 classroom teachers, drawn from the nearly 1.5 million teachers in the nation, was used for Teacher Opinion Poll IV conducted in the spring of 1962.

This sample was randomly divided into two groups. For an eight-week period, Group A received personalized treatment and Group B received nonpersonalized treatment. At the time the experiment was designed, we believed that the power of nonpersonalized letters to elicit responses was substantially less than that of personalized letters. Because of this belief, a safety factor was added to Group B to assure a high terminal response rate. At the end of eight weeks and at the end of ten weeks nonrespondents in Group B as well as Group A were to receive personalized letters.

Although the personalized treatment for Group B after eight weeks was included as a safety factor to assure useable opinion data, it was believed that any substantial closing of the response gap that existed between the two groups at the end of eight weeks and at the end of the experimental period would be additional evidence on the relative power of personal and impersonal treatments.

The treatment applied to the two groups are given in Table 2.

The contents of all communications was the same for both groups. The outside address of the envelope contained the name and school address of the individual teacher in both groups. All mailings were the same for both groups. All postage stamps were identical for both groups. The four factors which were varied during the first eight weeks are shown in Table 3.

Results of Personalized vs. Nonpersonalized

The experimental Hypothesis II was that personalized initial contact letters and follow-up procedures have the power to elicit a significantly higher response rate from teacher populations in mailed surveys than do nonpersonalized letters.

The statistical or null hypothesis was that there was no difference between the response pattern shown by Group A and the response pattern shown by Group B.

Figure II shows graphically the response for the two groups.

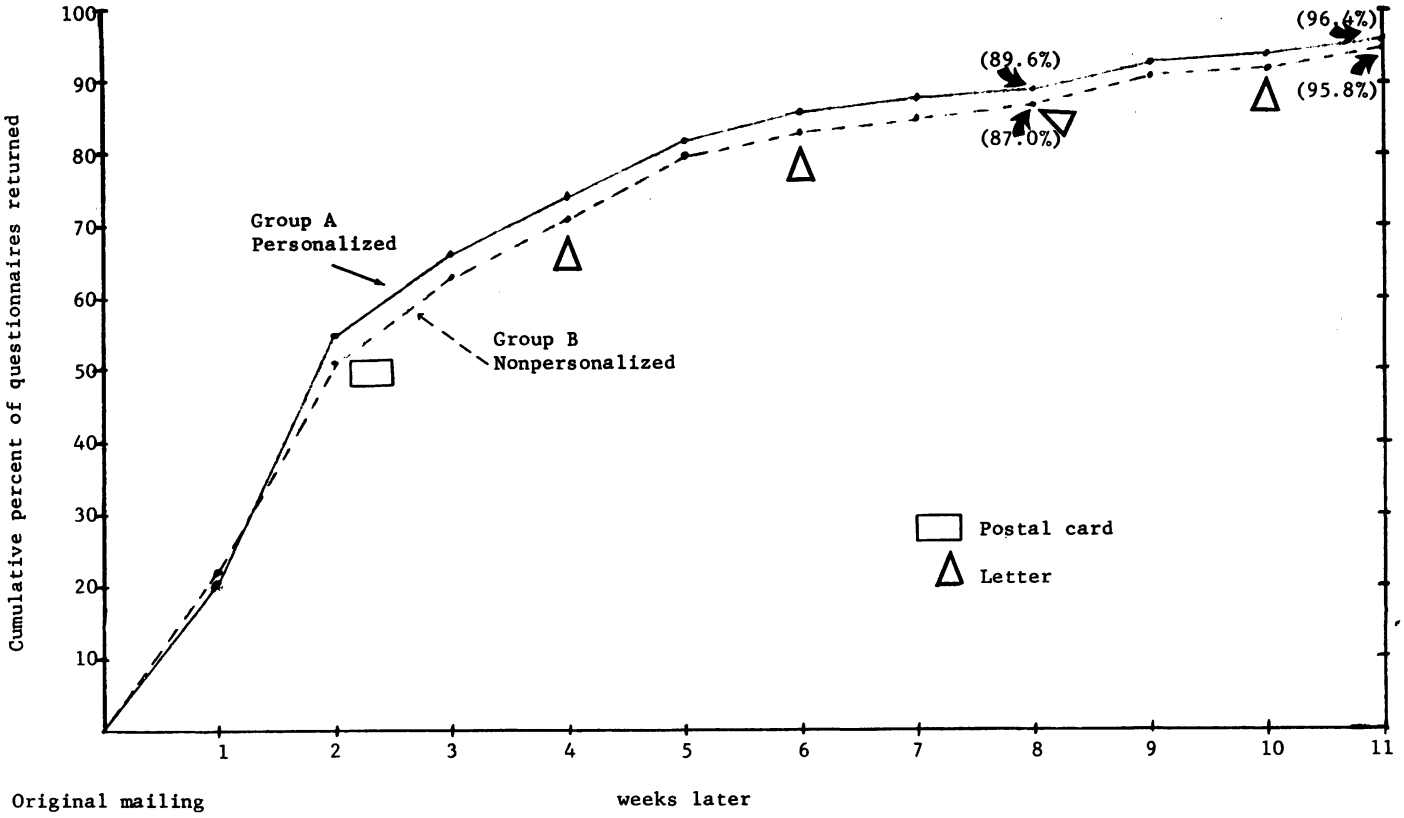
Table 4 presents cumulative weekly responses by number and percent for the two groups during the 11-week period.

A comparison of the response data for the experimental and control groups reveals the following:

TABLE 4.--CUMULATIVE RECEIPTS AND PERCENT OF RESPONSE FROM INITIAL MAILING AND SUCCESSIVE FOLLOW-UPS BY WEEKS FOR TEACHERS RECEIVING PERSONALIZED AND NONPERSONALIZED LETTERS, TEACHER OPINION POLL IV, GROUPS A AND B

Weeks from initial mailing	Group A (Experimental) n = 732		Group B (Control) n = 732		Percentage point difference in predicted direction (A > B)
	Number	Percent	Number	Percent	
1	2	3	4	5	6
1	153	20.9%	163	22.3%	-1.4
2	405	55.3	379	51.8	3.5
3	489	66.8	467	63.8	3.0
4	560	76.5	539	73.6	2.9
5	607	82.9	586	80.1	2.8
6	632	86.3	614	83.9	2.4
7	649	88.7	626	85.5	3.2
8	656	89.6	637	87.0	2.6
9	681	93.0	670	91.5	1.5
10	691	94.4	677	92.5	1.9
11	706	96.4	701	95.8	0.6

FIGURE II--CUMULATIVE RESPONSES FROM TEACHERS RECEIVING PERSONALIZED AND NONPERSONALIZED INITIAL AND FOLLOW-UP LETTERS IN TEACHER OPINION POLL IV, GROUPS A AND B



1. With the exception of the first week, the response rate for the experimental group consistently exceeded that of the control group. This factor would suggest that the personalized treatment was affecting response in the predicted direction.

2. However, these differences were surprisingly small, ranging from 2.4 to 3.5 percentage points in the predicted direction during the initial 8-week period. The average percentage point difference was 2.9 for the eight weeks. When these differences were subjected to the Kolomogorov-Smirnov test, it was shown that differences this great could be expected to occur by chance alone at well above the .10 level. Thus, there is not sufficient evidence to cause us to reject the null hypothesis.

3. Therefore, we must reject, with reservations, experimental Hypothesis II and conclude that in studies of this type personalized letters do not significantly affect the response rates exhibited by samples of teacher populations. In reaching this conclusion, there are several conditioning factors which will be discussed later.

4. At the end of the eighth week, the difference between the experimental and the control groups was 2.6 percentage points in the predicted direction. At this point the control group began to receive the same personal treatment as the experimental group. Within the three weeks that followed, the percentage point gap between the control and the experimental groups was closed to the point that the control group lagged by only 0.6 of a percentage point. Although these observations are interesting and are consistent with the slight difference shown earlier, the difference is so small that it is attributable to chance.

Some General Conclusions and Observations

From our experimental work in the field of sample surveys mailed to teacher populations, we draw the following general conclusions and make these observations:

1. One of the criticisms frequently directed at those who use mailed survey techniques is the generally poor response. The NEA Research Division's Sampling Project has overcome this criticism through its demonstration that consistently high response rates can be obtained from probability samples of teacher populations.

2. Our experience demonstrates that high response rates (above 90%) can be obtained at a reasonable cost. We estimate the costs for mailing and typing the personalized initial and follow-up letters described for Group A to be $30 per 100 persons in the sample. We estimate the cost of mailing and duplicating of the nonpersonalized letters and the two personalized letters described for Group B to be $17 per 100 persons in the sample.

3. We find no evidence that the use of telegrams as a medium for follow-up communications with teacher samples increases the final response rates where time is not a factor.

4. We find evidence that the use of telegrams tends to hasten the response of teacher samples.

5. In this study the differences between the response pattern of the samples of teachers receiving personalized and nonpersonalized letters was not statistically significant.

Similar findings were reported by Clausen and Ford[3/] in studies of the responses of World War II veterans to mailed surveys on attitudes toward, and information about, National Service Life Insurance. They reported:

> In mail follow-ups of veterans who had not responded to the initial questionnaire, personalized salutation and true signature did not lead to significant increases over nonpersonalized forms in rate of response....

Although the findings of our study failed to show a significant difference in the response patterns as between personalized and nonpersonalized approaches, we would caution against generalizing to surveys that differ either in type of population studied or in nature of survey instrument used.

The survey instruments used in these studies were short, multiphasic questionnaires. Their content covered several subjects of great interest to teachers generally. Although there is evidence from our experience and from the experience of other researchers that the length of the questionnaires does not have a significant effect on response rates, there is evidence that multiphasic questionnaires tend to have a greater drawing power for response than do monophasic questionnaires.

Clausen and Ford[4/] found:

> A multiphasic survey, covering several potentially interesting topics, yielded higher rates of response than a single subject survey of the same population, and also greatly lessened an interest bias in response.

Quite possibly the reason for the low differential in drawing power between personalized and nonpersonalized letters found in our study is an overriding high-interest effect. It could be reasoned that in a study in which the survey instrument is monophasic and confined to a single area of inquiry of low interest to teachers, the superior drawing power of personalized letters might be significantly demonstrated. In studies presently being planned, we hope to test this possibility.

3/ Clausen, John A., and Ford, Robert N. "Controlling Bias in Mail Questionnaires." Journal of the American Statistical Association 42: 497-511; December 1947.

4/ Ibid., p. 497.

6. It is our tentative conclusion from our experience to date that an optimum contact and follow-up procedure where cost-quality relationships are concerned and where time is not a crucial factor, consists of a combination of the personalized and nonpersonalized approaches. Such a procedure might consist of a personalized initial letter, nonpersonalized postal card reminder, nonpersonalized letter, and a personalized letter sequence. In our opinion the optimum time between contacts seems to be one to two weeks. We are planning studies to test the validity of these assumptions.

A NEW METHOD OF PROBABILITY SAMPLING IN THE FIELD OF EDUCATION (Abstract)

Philip Desind, U. S. Office of Education

In this paper is described a method for converting a two stage sample plan to a one stage design. At the same time the proposed method which is based upon the frequency of the letters of last names, permits maximization of the primary units for a given size sample of secondary units. This method is highly advantageous since the administration of setting up the design by the respondents is simple.

XIII

RESEARCH ON HUMAN AGING

Chairman, Anita K. Bahn, National Institute of Mental Health

Page

A Longitudinal Multidisciplinary Study of Human Aging: Selected Methodological Issues - George L. Maddox, Duke University..... 280

The Fate of a Panel Over a Five-Year Period - Wayne Wheeler, University of Chicago..... 286

Intellectual Change in the Senium - Robert W. Kleemeier, Washington University..... 290

Discussion - Samuel W. Greenhouse, National Institute of Mental Health..... 296

Discussion - Marvin Taves, U. S. Department of Health, Education, and Welfare..... 298

A LONGITUDINAL MULTIDISCIPLINARY STUDY OF HUMAN AGING: SELECTED METHODOLOGICAL ISSUES

George L. Maddox, Duke University

The Department of Psychiatry at the Duke University Medical Center is conducting an exploratory study of human aging focused on selected physical, psychological, and social correlates of the aging process. In this study a number of the aging processes have been studied simultaneously by persons representing various disciplines. The intention has been to maximize the number of variables which might be introduced in the study of any single process of change as well as to explore the interrelationship of various processes.

Aspects of the design for this research will be described, and some of their consequences for the interpretation of data will be assessed. Two characteristics of the research design are of special relevance for this discussion:

1. The subjects of the investigation are volunteers from among persons 60 years of age and over who were living in a community at the time they indicated a willingness to come to the medical center for two days of interviews and clinical evaluations.

2. The focus of the research is on the processes of change as well as on the changes commonly observed among elderly subjects. Continued observation of the same subjects over an extended period of time has been considered basic.

Sampling procedure and panel maintenance in longitudinal studies are not novel methodological issues. They are, however, persistent issues, and the method of handling them is fundamentally related to the analysis, interpretation, and generalization of research data. The purpose of this paper is to illustrate this relationship within the context of a specific research enterprise. In order to place decisions about research design in context, it will be helpful to describe briefly the development of this particular research project. The consequences of these decisions for data analysis will then be assessed. No attempt will be made to summarize substantive findings.

Origins of the Research Project

The research under consideration is generically related to an investigation of senile patients begun more than a decade ago at the University of Colorado School of Medicine by Dr. E. W. Busse and his associates. Initially this research was focused on changes in the aging central nervous system as revealed by electroencephalographic records. In order to develop standards for evaluating the clinical records of senile elderly persons, volunteer subjects with a wide range of socioeconomic characteristics were used as normal controls. The volunteers included the old and the young, indigents and professional persons, the employed and the unemployed, the sick and the well. The initial findings suggested differences between the EEG records of institutionalized and non-institutionalized subjects; moreover, among non-institutionalized subjects, differences were related to variations in health and in socioeconomic circumstances.

The Colorado experience demonstrated the importance of extending the use of normal controls in developing standards for the interpretation of EEG records in elderly subjects. It also suggested the need to investigate interrelationships among the physical, psychological, and social correlates of these differences, and changes in their relationships through time.

The research was subsequently continued at Duke, and by 1957 electroencephalographic research was being conducted simultaneously among elderly persons in a private home for the aged, senile patients in a state mental institution, and non-institutionalized subjects living in and around Durham, North Carolina. These last subjects, recruited individually and through various community organizations serving older persons, included retired professionals and executives as well as persons from lower socioeconomic strata.

Other medical specialists and social scientists joined the psychiatrists in making plans for extending the scope of the research to explore the correlates of central nervous system change. In this planning, understanding differences in the changes experienced by the same elderly person through time emerged as the central issue. Initial interest in the non-institutionalized subject developed as a reaction against the obvious inadequacy of applying observations made on institutionalized elderly subjects to elderly persons in general. This interest was given a decided boost by increasing national concern about the problems of non-institutionalized elderly persons. Consequently, it was decided to undertake a longitudinal, multidisciplinary investigation of the aging central nervous system and its correlates among non-institutionalized elderly subjects.

Selection of Subjects

The issue of sampling appropriate for the proposed research posed a basic problem. A review of the minutes of the seminar within which the research design was developed indicates that the discussion went something like this:

Persons engaged in a scientific enterprise can concern themselves with generalizations about the distribution of some characteristic or variable within a population; the epidemiologist,

for example, is preoccupied with this type of generalization. But a scientist may also be interested in the relationship between or among characteristics or variables independent of their distribution in a population. The scientific experiment that seeks to relate variables X and Y under given conditions is a case in point. Generalizations of the first type require the investigation of some universe or of samples representative of that universe. Generalizations of the second type require, as a minimum, control of variables presumed to be relevant. While optimum research design should provide for both types of generalizations, practical considerations frequently make it necessary to choose one or the other type as an immediate goal.

With appropriate misgivings, the research group at Duke decided to continue the practice of recruiting volunteers rather than attempting to involve a randomly drawn sample of elderly non-institutionalized subjects in research that would require intensive clinical evaluation. In recruiting more volunteers from the community however, an attempt was to be made to develop a panel of subjects 60 years of age and older whose age, sex, ethnic, and socioeconomic characteristics would reflect the range of these characteristics found among older persons in the locale. It was not the purpose of this procedure to provide the appearance of random sampling without its substance, but rather to maximize the variety of controls which could be introduced subsequently. In time, 250 subjects who met these criteria indicated a willingness to participate in the research and completed the initial two-day series of interviews and clinical examinations. The only inducements offered these volunteers were a free annual medical examination, a subscription to a magazine for elderly persons, and whatever satisfaction the subjects derived from their participation.

One consequence of the decision to use volunteer subjects was obvious. Generalizations from the data would necessarily be limited to statements concerning the relationship among factors in the aging process under specified conditions. Statements about elderly persons in general would not be warranted.

The deliberate decision of a research group to use volunteers and to accept the consequences of its decision does not have to be defended. But an interesting question is posed: Was there in fact a feasible alternative to their decision? When the research project is one which necessarily involves clinical evaluations and which must be carried out among non-institutionalized subjects, the answer seems to be no. Attempts to involve randomly selected non-institutionalized subjects in research requiring clinical evaluation have not proved rewarding. It is difficult to get subjects living in the community into a clinic for a single examination, much less for a series of different examinations; and the problem is intensified with the increasing age of the subjects.[1]

The decision to use a panel of volunteers raises additional problems concerning subject selection which are not resolved by restraint in the generalization of findings. In addition to being a crucial condition of generalizing from a part to a whole, random selection procedure also has another function in research: It presumably minimizes the probability of introducing systematic bias from unrecognized or uncontrolled but potentially significant sources into the analysis of relationships among variables. For example, volunteers used as "normal controls" in clinical research may in fact neither be normal nor provide controls. In specific instances it has been demonstrated that volunteer subjects have introduced an uncontrolled but definitely biasing effect in clinical experiments.[2] The fact that an individual is not institutionalized and not a patient does not necessarily make him a normal control within any frame of reference.

Subtle as well as gross biasing effects introduced by self-selection may also have a bearing on the analysis of relationships among variables. The Duke research project, for example, was designed to explore selected physiological, psychological, and sociological correlates of central nervous system aging. Each type of factor implies a number of universes within which subjects may be ordered in some fashion -- say, in terms of relative health or illness, of higher or lower intelligence, of greater or lesser age, or of higher or lower socioeconomic status. Since a given individual is a member of each of these universes simultaneously, it is difficult to parcel out the effect of each factor in a given situation. If the subjects are volunteers, they may represent adequately the theoretically possible range of health and illness but include only those who are of higher intelligence and upper socioeconomic status; or they may run the gamut from high to low socioeconomic status but include only those who are relatively healthy and of high intelligence; or the subjects may be concentrated among those who are in good health, have high intelligence, and are in the upper range of status.

In the absence of probability sampling, the possibility of spurious relationships which are artifacts of the sampling procedure is increased. A cross sectional analysis of project data, for example, indicated that activity, but not morale, decreased with age among these elderly subjects. That this finding is probably an artifact of the age distribution of antecedent or intervening variables known to affect the relationship between activity and morale will be illustrated later.

A number of studies involving non-institutionalized elderly subjects have employed random sampling procedure. The non-participation rates in such studies have been high, especially when clinical evaluation of subjects has been involved. It is entirely possible that a high refusal rate in such instances produces a sample which is essentially the equivalent of a collection of volunteers. To investigate this possibility, the Duke panelists have been compared with samples of elderly persons used in other research projects.

Three types of comparisons have been made: (1) a comparison in terms of selected physical and mental health characteristics between the Duke panelists and samples presumed to be representative of local, regional, or national populations of older persons; (2) a comparison of the panel with itself at two points in time, on the assumption that selective dropout of subjects through time might provide a rough indication of the characteristics of subjects initially attracted to participate in the research; and (3) a comparison, in terms of selected characteristics, between some of the panelists and a probability sample of elderly white subjects drawn from the same community by another research group for a survey of community involvement and participation.

Selected comparisons with randomly drawn, non-local elderly subjects

Physical health and associated degrees of disability are principal variables in any investigation of the aging central nervous system. The Duke research excluded the minority of elderly persons totally confined to their home by physical disability. Whether, among the majority of mobile persons 60 years of age and over, those in poor or good health would be more likely to participate is problematical. A partial answer to this question is provided by a limited comparison with certain findings of the National Health Survey.[3]

The comparison is limited by the fact that the National Health Survey utilized the reports of respondents as the basis for assessing degrees of disability, while the comparable health evaluation of the Duke panelists was based on clinical examinations. Moreover, pathological conditions appear characteristically to be under-reported in surveys.[4] One comparison which does seem warranted, however, involves the category of persons in each study rated as severely disabled. The probability of coincidence between self-reports and clinical evaluations should be maximum in the extreme cases. Thirty-three per cent of the persons 65 years of age and older who were interviewed in the National Health Survey reported that they were either restricted in major life activities or totally confined to the home as a result of poor health. On the initial clinical examinations, 26 per cent of the Duke panelists were assessed to be at least 60 per cent disabled in carrying out normal life activities; and, when the panelists were re-evaluated approximately three years later, 33 per cent were so disabled.

The presence or absence of physical pathology is related to an individual's assessment of his health status, but is not synonymous with it. Self-assessment of health status thus provides another basis for comparing the Duke panelists with a probability sample. Schnore and Cowhig[5] have reported the distribution of self-assessment of health status among a randomly drawn sample of persons living in several metropolitan centers. Among their respondents 60 years of age and over, 46 per cent assessed their health as "excellent" or "good"; among Duke panelists, 52 per cent made this assessment.

Neither of the foregoing comparisons shows a striking difference between the volunteer panelists and randomly selected subjects who were willing to participate in other studies.

A serious objection advanced against the use of volunteers in clinical research is the large proportion of persons with diagnosed or diagnosable psychiatric illness who tend to be attracted.[6] Among the panelists at Duke, however, 40 per cent were found to be without significant signs or symptoms of psychiatric illness, and only 6 per cent were classified as psychotic. This distribution between the extremes of mental health and illness is similar to comparable data based on mental health surveys utilizing random sampling techniques.

An example is Gruenberg's epidemiological study of mental illness among elderly subjects in selected census tracts in Syracuse.[7] Approximately 4 per cent of the respondents were classified by him as dangerous to themselves or unable to care for themselves. While the home survey sample of Srole and his associates in their study of mental illness in a metropolis[8] does not include persons beyond the age of 59, only 38 per cent of the subjects between 50 and 59 years of age were considered to be either "well" (unimpaired) or to exhibit only "mild" symptoms. Dr. Leighton, in an epidemiological survey of mental health in a small town in Nova Scotia, also found that 14 per cent of the respondents aged 18 and over were "asymptomatic" and another 21 per cent were borderline "normals."[9] Although she did not analyze her findings in terms of specific age categories, the Duke panelists obviously compare faborably.

Selected comparisons of the physical and psychiatric characteristics of the Duke volunteers with subjects selected by random procedures do not indicate striking differences. On the whole, the physical and mental health of the Duke subjects appears to be slightly better than that of the samples of elderly persons with whom they are compared. These similarities do not prove that the Duke subjects provide the

equivalent of a randomly drawn sample of elderly subjects; rather, they indicate that the relatively high refusal rate among older persons asked to participate in surveys tends to make the participating respondents essentially the equivalent of volunteers.

Selective dropout among panelists

Of the 250 subjects for whom there was relatively complete information at the end of the first phase of the study, 182 (73 per cent) repeated the two days of interviews and examinations approximately three years later. Of the 68 who failed to return, 65 per cent had died in the interim or had been immobilized by illness, 11 per cent had moved away, and the remaining 24 per cent indicated in various ways a reluctance to continue participation.

Three characteristics were observed with significantly greater frequency in the subjects who refused to return for the second phase of the research than in those lost because of death or immobility. Refusals were concentrated among (1) females, (2) subjects with IQ (WAIS) scores below the median of the panel, and (3) those clinically assessed to be in relatively good health (less than 20 per cent disability). In the case of this last factor, the voluntary dropouts were concentrated among subjects who were medically assessed to be in good health, but who assessed their own health status as poor. Subjects who refused to return also tended to be of lower socioeconomic status and relatively inactive socially, although these tendencies were not statistically significant.

When the 182 panelists who returned for the second evaluation are compared with the 250 original members of the panel, a significant increase is found in the proportion of subjects characterized as (1) active, (2) married and living with the spouse, (3) the head of a household, (4) in good health, (5) exhibiting high morale, and (6) intelligent (above median WAIS scores). The proportion of Negroes, females, and persons of higher social status among the panelists also increased, although these increases were not statistically significant.

Over the three years under consideration, the panelists as a group tended to become increasingly a physiological, psychological, and social elite. While part of this change reflects the effect of selective mortality, it is the impression of various members of the Duke research group that the volunteers initially represented a relatively elite category of elderly subjects, and only became more so through time. Supporting this impression are (1) the comparisons of the panelists, in terms of physical and mental health, with randomly drawn samples of elderly subjects who have "volunteered" by not refusing to cooperate in other studies; and (2) the observation that cooperation in a strenuous two-day sequence of interviews and examinations tends to exclude the physically disabled, the psychologically deteriorated, and the most socially deprived elderly persons in the community.

Analysis of changes in the composition of the panel through selective dropout does nothing more than suggest differences between the panelists and the population from which they came. The noted changes do, however, have important implications for data analysis in longitudinal research. These will be discussed later.

Comparison with a random sample of elderly subjects in the same locale

For a study of community participation among white persons of various ages, the Department of Sociology at Duke drew two area probability samples in Durham -- one including persons through the age of 64, and another including persons 65 years of age and older. If the original panelists who are 1) Negro, 2) under 65, and 3) living outside the city limits are excluded, it is possible to compare those remaining (N=108) with those persons 65 years of age and older in the probability sample (N=139) in terms of age, sex, marital status, activity, and self-estimate of health.

In terms of age and self-estimate of health, the probability sample of persons 65 years of age and older and the comparable panelists are similar. The average age of the sample subjects was 72, and that of the panelists, 71; 14 per cent of the former and 15 per cent of the latter were 80 years of age or older. Sixty-two per cent of the panelists being compared estimated their health to be good or excellent, while 56 per cent of the sample subjects made this estimate.

In terms of reported activity, sex distribution, and marital status, however, these two groups of elderly subjects showed striking differences. The mean Activity Inventory score[10] originally of the Duke panelists was 27.4, as compared with 20.4 for the subjects in the probability sample. Moreover, 54 per cent (compare 36 per cent) of the panelists were male, and 65 per cent (compare 34 per cent) were married and living with the spouse. The high proportion of panelists who were married and living with their spouses can be explained by two factors: (1) the high proportion of males in the group, and (2) the fact that there were 13 married couples participating. That even one couple would be drawn in a random sample of individuals is extremely unlikely. Since "married and living with spouse" is a factor positively associated with physical and mental health, the high proportion of persons in the panel who fall into this category helps to ex-

plain the favorable showing, in terms of health, made by the panelists in comparison with other groups.

Summing up, this comparison with a random sample of elderly subjects in the same community tends to support the inference that the Duke panelists represent an elite among the elderly. It is all the more interesting, therefore, that the panelists are so nearly similar, in selected health characteristics, to randomly drawn samples of elderly subjects in the instances noted.

A Longitudinal Design

The initial conception of the Duke research project was to explore human aging as a complex process of interrelated changes. Selected physical, psychic, and social factors involved in this process were to be explored. A longitudinal design was proposed because it permitted intensive focus on individual subjects and also permitted each subject to be used as his own control.

Baselines for each subject were established in the first series of interviews and examinations. A cross-sectional analysis of the initial data explored the complex interrelationships among a wide range of variables and focused attention on those factors which would warrant the most attention in the longitudinal analysis. The second series of observations on the subjects who remained approximately three years later provided a basis for assessing individual changes and an additional check on the relationships initially observed.

The cross-sectional analysis proved to be a double temptation:

1. In addition to providing information about the relationship between factors under specified conditions, such an analysis also provides information about the distribution of these factors among the panelists and, in some cases, measures of central tendency.

Even though the temptation to draw unwarranted inferences about elderly persons from such distributions and statistics might be avoided by an investigator, there is little control over the inferences which others draw. By a conscious focus on processes of aging and their correlates among selected older persons rather than on the description of the population of elderly persons, the Duke investigators have kept such errors to a minimum.

2. A panel statistic based on observations at two different points in time understandably invites comparison. That such comparisons must be made with great caution when there is a selective drop-out of panelists is suggested by the following illustration:

Both the initial and later cross-sectional analyses indicated a significant positive relationship between contact with the environment (activity) and morale among the panelists. It was hypothesized that, as activity tends to decrease with age, so should morale. When the age factor was controlled, however, both initial and later analyses showed an age-related decrease in activity, but not in morale. In the later analysis, moreover, the relationship between the activity and morale, while still significant, was less pronounced. A possible explanation for this observation is the "theory of disengagement" suggested by Cumming and Henry on the basis of their Kansas City data.[11] These authors hypothesize that, with age, activity and morale become increasingly independent of each other.

Analysis of the Duke data suggested an alternative interpretation. The initial cross-sectional analysis disclosed three important factors in interpreting the relationship between activity and morale: (1) health, especially an individual's own assessment of his health status; (2) the presence or absence of clinical depression; and (3) the individual's reported sense of usefulness. Morale was likely to be higher in elderly panelists who were, or believed themselves to be, in good health, who were not depressed, and who felt useful than in those who reported an equal degree of activity, but did not have these other characteristics. Moreover, the panelists who showed these characteristics were more likely than the others to have high morale, even though they reported low activity.

When the distribution of these modifying characteristics was checked against age, it was found that the proportion of subjects with one or more of these characteristics was slightly higher in the older age categories than among the younger panelists. This difference was even more noticeable when the later series of observations were made. In this instance, the finding that morale became increasingly independent of activity with age appears to have been an artifact of the particular characteristics of the subjects in the various age categories. The slight decrease in the mean activity score of the panel over approximately three years, concomitant with an increase in the mean morale score, can be explained in the same way. When individual subjects are used as their own controls, changes in activity scores tended to be related positively to changes in morale <u>except</u> in the presence of one or more of the factors known to modify the relationship.[12]

This illustration re-emphasizes the soundness of the initial orientation of the research group to use each individual as his own control and, when comparisons of panelists in terms of group statistics seemed appropriate, to match the comparison groups with care. It also serves as a reminder that, when summarizing

measures are used in longitudinal studies, an observed relationship may be an artifact of changes in the composition of the panel resulting from selective dropout.

Summary

Selected aspects of a research design have been described, and some consequences of this design for data interpretation have been discussed. The research project involved a longitudinal multidisciplinary investigation of the aging process and its correlates among selected non-institutionalized subjects 60 years of age and older. Two hundred and fifty elderly subjects volunteered to participate initially in a two-day series of interviews and clinical examinations. Approximately three years later, 182 of them returned to complete a second series of investigations.

Two methodological issues discussed on the basis of experience in this investigation are: (1) the use of volunteer subjects and (2) selective dropout of panelists in a longitudinal study. Regarding the first, it is argued that, in spite of the limitations incurred, there is at present no feasible alternative to the use of volunteers in research requiring clinical examinations. Selective dropout of subjects imposes still further limitations on the use of group statistics in the analysis of longitudinal observations. Because of the changes in the characteristics of groups under investigation over a period of years, such statistics are dependable only if there is control for the effects of selective dropout of subjects.

FOOTNOTES

1. See, for example, S. Cobb et al, "Differences between Respondents and Non-respondents in a Morbidity Survey Involving Clinical Examination," Journal of Chronic Disease, 6 (1957), 95-108; Tava Gordon et al, "Some Methodological Problems in the long-term Study of Cardiovascular Disease: Observations on the Framingham Study," Journal of Chronic Disease, 10 (1959), 186-206; P. Borsky and E. White, "Factors Affecting the Decision to Volunteer for a Physical Examination," Public Opinion Quarterly, 23 (Fall, 1959), 445-446; "Cooperation in Health Examination Surveys," United States National Health Survey (Publication D-2), June, 1960; and D. E. Krueger, "Measurement of Prevalence of Chronic Disease by Household Interviews and Clinical Evaluation," American Journal of Public Health, 47 (August, 1957), 953-960.

2. H. Esecover et al., "Clinical Profiles of Paid Normal Subjects Volunteering for Hallucinogen Drug Studies," American Journal of Psychiatry, 117 (1961), 910-915, presents a case in point and a convenient bibliography. It is pertinent that most discussions of volunteer "normal controls" concern only younger subjects. Experience in the research reported here suggests that the usual characterization of the young volunteer as "not normal" cannot be extended uncritically to apply to elderly subjects.

3. U. S. National Health Survey, "Preliminary Report on Disability, United States, July-September, 1957" (Series B-4).

4. Krueger, op. cit.

5. "Some Correlates of Reported Health in Metropolitan Areas," Social Problems 7 (Winter, 1959-60), 218-225.

6. Esecover et al., op. cit.

7. "A Mental Health Survey of Older Persons," in Hoch and Zubin (Eds.), Comparative Epidemiology of Mental Disorders (Grune and Stratton, 1961).

8. Mental Health in Metropolis, I (McGraw-Hill, 1962).

9. "The Distribution of Psychiatric Symptoms in a Small Town," American Journal of Psychiatry, 112 (1956).

10. See R. S. Cavan et al., Personal Adjustment in Old Age (Chicago: Science Research Associates, Inc., 1949).

11. Growing Old (Basic Books, 1961).

12. For a more detailed discussion of this point, see the author's "Activity and Morale: A Longitudinal Study of Selected Elderly Subjects," a paper presented at the annual meeting of the American Sociological Association, Washington, D. C., September, 1962.

THE FATE OF A PANEL OVER A FIVE-YEAR PERIOD

Wayne Wheeler, University of Chicago

This year marks the tenth anniversary of the studies of adult life carried out in Kansas City by the University of Chicago's Committee on Human Development. "Adult life" is a euphemism for age and aging. The studies have dealt primarily with personal and social adjustment associated with middle and old age. They have been also concerned with the sociological and psychological processes of change which are usually termed "aging."

My purpose here today is not, however, to discuss any substantive findings of the Kansas City research. It is to describe the history and fate of the longitudinal panel from which the data were collected in a series of interviews carried out at intervals over 5½ years from September, 1956, through March, 1962.[1]

The major goal of the research with this panel was to acquire, through time, a rich source of information and insight from which hypotheses concerning the aging process could be developed. It was felt that continued contact with the respondents would allow the collection of data serially, the augmentation of information about each case, and the addition of new questions as knowledge increased. For these reasons and because repeated contacts with respondents over a five-year period could easily produce a mountain of unmanageable data, the size of the panel was kept small.

The "base sample" from which the panel for the study of adult life was drawn, consisted of 8,300 households and 400 persons in quasi-households selected randomly from the Kansas City Metropolitan Area. These 8700 dwelling units comprised over 95% of the original area sample.

Time does not permit a complete discussion of the sample which came ultimately to make up the study's panel. Briefly, however, the details are as follows: The study design called for a sample of respondents that was stratified relative to three variables: age, sex, and social class. Further, the sample was to include non-institutionalized white persons between the ages of 49 and 71 who were not unable, because of serious disability, to perform in their usual capacities such as work or housekeeping. Respondents were also to be restricted to those living in the Urbanized Area as contrasted to the Metropolitan Area which also included farm and village inhabitants. This Urbanized Area includes both Kansas City, Missouri, and Kansas City, Kansas, and contigious suburbs. It contains approximately 95% of the population of the Metropolitan Area. After elimination of the cases which did not meet the established criteria, there remained 1236 cases to score for social class. This was done by using an adaptation of the Warner Index of Social Characteristics carefully worked out[2] over a period of years for Kansas City, and by using data gathered in a survey of cases from which the panel was drawn. On the basis of ISC scores, upper- and lower-lower class persons were eliminated from the 1236 cases.

The cards, each representing one of the remaining cases, were then sorted into age, sex, and social class categories, assigned numbers, and through use of a table of random numbers, cases were chosen with which Wave I of the interviewing was to be attempted.

Interviewers were given a total of 236 assignments on Wave I. Of these, 88 males and 86 females completed the interview and agreed to cooperate with the research. Thirty-seven potential respondents refused, and 25 potential respondents were not available for study because they had moved, died, or were unobtainable for other reasons after data on the base sample had been gathered.[3]

Before going into my discussion of the fate of this main panel, the selection of which I have just described, I should like to mention another portion of the study population from which we have obtained data. This is a supplementary quasi-sample made up of persons between the ages of 70 to 90. It was added at the time of Wave III of the main panel in order to give a greater age span than that represented by the main panel itself. It covered the period from June, 1957, through March, 1962. Difficulty and expense in obtaining a random sample of healthy, very old persons would, of course, be created by the scarcity of such people in the general population. This supplementary sub-grouping was, therefore, recruited into the study population from among neighbors, acquaintances, and other sources of three interviewers, one each from the upper-middle, lower-middle, and upper-lower classes. Social class status was assigned to these very aged respondents on the basis of information gathered during interview Waves III through VI.

The compositions of the panel at the completion of Wave I and the quasi-sample at the completion of Wave III will be seen in Tables Ia and Ib and Tables Ic and Id, respectively.[4] From these tables we also get a picture of the fate of the main panel through the seven successive waves of interviewing and of the quasi-sample through five waves.

In Tables Ia through Id, commencing with Wave II, a sub-category of completed interviews termed "pick-ups" is introduced. A "pick-up" is a respondent marked as a refusal after persistent attempts to interview during the period of a particular wave but who at a later date--in most cases after the completion of Wave VI--consented to be interviewed one or more times. A pick-up may or may not be a reinstatement into the study because a completed interview associated with a particular wave does not necessarily indicate the respondent consented to interviews associated with other waves. Conversely, the same respondent might be listed as a pick-up on several interviews.[5]

The number of attritions or drop-outs due to refusals, deaths, or moves which could not be traced or were away from the Kansas City area, is also recorded in Tables Ia through Id commencing with Wave II. A number of devices were employed in an effort to keep the attrition rate at a minimum. The study was explained in considerable detail to each respondent, and his cooperation for a five-year period sought at the time of the

first interview. Hard-sell methods were avoided because it was felt that sustained cooperation could not be obtained by pressure tactics. Each respondent received a greeting card from the project director during the holiday season.[6] Upon the completion of an interview, the respondent received a thank you note from the project director.

In general, it was felt that once an interviewer established rapport with a respondent, that interviewer should continue to be assigned the respondent. Changes in the interviewing staff did not always make this possible, and for Wave VII it was decided that no respondent should be interviewed by a staff member by whom he had ever been previously interviewed. Further, if a respondent and an interviewer did not hit it off, another interviewer was given the assignment.

After Wave VI, when contact with special study groupings, to be discussed later increased, feedback indicated that a sizable proportion of respondents were becoming restless because of the demands being made upon them. It, therefore, seemed especially desirable to attempt to increase the sense of involvement of respondents in the project. To that end, a number of personal letters--some associated with specific interviews, some not--were sent to each respondent from the project director. A pamphlet explaining the research in lay terms was distributed freely among respondents and their significant others during an investigation of small social systems carried out between Waves VI and VII. Also, after Wave VI, an especially competent, highly trained interviewer was given the task of attempting interviews with persons who had been marked "final refusals" sometime during the course of the research. Nearly all of the previously mentioned "pick-ups" were moved from the refusal category to the completion category through the efforts of this interviewer.

From Tables Ia and Ib, we note that 67% of male panel members remained with the study at the end of 5½ years while 59% of female respondents on the panel remained. In each wave, with the exception of Wave IV, the attrition rate for females ran one or two per cent above that of males. This is inspite of higher death rates, as expected, for males. For both males and females, the range of attrition rates by social class and age category for each of the six waves after Wave I was from 1% to 10% with both the mean and the median at 6%. The rate of final refusals for female panel members for the whole study was 34% while for male panel members it was but 19%. Except for upper-lower class males in the 49-59 age grouping, the rate for females refusing to be interviewed was greater than that of males for all categories. Out of 174 persons completing Wave I, eight--four males and four females--were lost through moves. The females, who moved were, however, concentrated in the lower-middle class while males who moved were from both the upper-middle-and the upper-lower, but not the lower-middle, classes.

Table IIa summarizes the total of combined male and female panel members by age category and social class who completed interviews during each of the seven waves and the per cent of the total of Wave I who completed Wave VII. Table IIb gives another perspective on the combined total, showing the total number of interviews completed by age, sex, and social class for each of the seven waves and the per cent of the total of Wave I who completed Wave VII.

The difference in rate of attrition between social classes was significant at the .05 level. The lower-middle class had higher attrition rates than either the upper-middle or the upper-lower classes. The reasons for attrition, i.e., refusals, deaths, moves, were individually evaluated, in terms of age, sex, and class, for their effect on attrition; no significant relationships were established between each of these variables and attrition. A within class analysis of attrition resulted in a significant difference at the .05 level, by class between females. One concludes that the relationship between social class and attrition was due to lower-middle class females who die, move and refuse more often than either the upper-lower or upper-middle class females.[7]

Tables Ic and Id show the completed interviews and amount of attrition for male and for female respondents, respectively, in the quasi-sample after each interview wave, by age and social class. From among the 50 male respondents aged 69-90 who were incorporated into the study group at the time of Wave III, only 20, or 40%, completed Wave VII. This compares with the 38 female respondents, or 67%, out of 57 in the same age categories and social class range who completed the same wave. Males had a higher percentage of total attrition in each wave. This difference between males and females is significant at the .05 level.

Death rates for males in the quasi-sample again ran higher than those for females, the relative percentages for the 3 3/4 year period being 26% and 18%. At the .05 level there is a significant difference in the death rates in each interview wave for males. Also, with regard to deaths, there is a significant difference between persons aged 69-79 and persons 80 years of age and over.

Table IIc shows the total of combined male and female quasi-sample respondents by age category and social class who completed interviews during each of the five waves and the per cent of the total of Wave III who completed Wave VII. Table IId shows the total number of interviews completed by age, sex, and social class for each of the five waves and the per cent of the total of Wave III who completed Wave VII.

It is worth noting that the proportion of respondents who were brought back into the study after they had been considered to be refusals is much greater for the quasi-sample than it is for the main panel, and indeed, this difference is significant. However, in contrast to the younger respondents making up the main panel, the final refusal rate for males in the quasi-sample was higher than that for females.

Lest some of my hearers get the impression that persons who did not abort from the panel were approached for interviews but seven times--five in the instance of the quasi-sample--, I would like to mention once again the special study groupings. The numbers of male and female respondents in both the main panel and the quasi-sample in these special groupings are shown in

Tables IIIa through IIId.

In all, interviews for five studies subsidiary to the general study were carried out. In general, the special interviews concerned with morale, affect-control, and kinship were designed to develop insights and hypotheses which could be tested in regular interview waves on the total study population. The life satisfaction interview, clinical in nature, was for the purpose of validating a scale, items for which were contained in several of the regular interviews.

The study of small social systems involved not only respondents, in the sense that names and permissions were obtained from them, but also interviews with their relatives, friends, and professional and business acquaintances whom we call "significant others." Formal interviews were carried out with respondents and with their "significant others" in the studies concerned with social and cultural systems in which respondents have their beings. Additional data on systems were gathered by two participant observers who lived in areas where they had the opportunity to make informal, continuous contact with several respondents over a period of up to six months. These contacts resulted in ethnographic-type reports. The last column in Tables IIIa through IIId indicates that this manner of approach to certain respondents was not entirely successful while it had considerable success with others.

A further word about length and number of interviews seems to be in order. Throughout the research, the minimum time required to complete an interview with any one respondent was in the neighborhood of one hour. The maximum occasionally reached six to eight hours, and three- or four-hour interviews were not unusual.

With respect to number of contacts with the study's staff, seven were, of course, the minimum for respondents in the main panel with whom the whole series was done. Five interviews were the minimum for respondents in the quasi-sample. It is well to remember, however, that the supplementary sub-studies added contacts. One woman included in the study of small social systems actually met with interviewers and other researchers 25 times in something over five years. Instances of 12 and 13 personal contacts between staff members and a respondent are not unheard of. Eight and nine interviews with main panel respondents are quite common.

In conclusion, let it be said that the longitudinal investigation of age and aging conducted in Kansas City from September, 1956, through March, 1962, involved a total of 281 respondents between the ages of 49 and 90. Of these, 113 or about 40%, aborted during the study. Of the 113, a total of 64 respondents or 37% dropped from the main panel of persons 49 to 71 years of age and 49 respondents or 46% dropped out of the quasi-sample of 69 to 90 year-olds.

It is difficult to say, given the demands made upon the respondents, their ages, and Kansas City as the location for the Study, whether this is a good record. Nor, with the increased use of surveys for scientific and commercial purposes, could one hope for as much success at a future date with a population identical albeit more adept at defending its privacy.

We can, however, say that in the current research we have been flexible and sympathetic but persistant in our approach to our respondents. This probably shows in the large number of persons who put up with so much for so long.

FOOTNOTES

[1]Supported by the National Institute of Mental Health through Grant Number M-9082 to the Committee on Human Development of the University of Chicago, William E. Henry, Principal Investigator, with Robert W. Havighurst and Bernice L. Neugarten. I wish also to thank Stephen J. Miller, Helen Haubold, and Helen Henry for valuable assistance in this paper.

[2]By Warren A. Peterson, Richard Coleman, John C. Scott, Jr., and others.

[3]Every attempt was made to reach persons who refused as many times, each time with a different interviewer, as was considered appropriate. When these approaches, coupled with letters, still failed to produce cooperation, the person was classified as a definite refusal. These refusals became the object of a special sub-study inaugurated for the express purpose of finding out whether they differed from the persons who consented to be included in the study. A mock television interview was developed which replicated, briefly, some of the key issues of the first interview. An experienced and persistent interviewer, new to the respondents, was able to obtain this shorter interview with 22 of the refusal group. Examination of these interviews has allowed us to compare them with the completions. In this respect, there are but 15 persons in the 236 persons on which little information is available.

[4]Space limitations preclude the inclusion of the lengthy tables accompanying this paper in these Proceedings. Persons who desire a set of these tables may obtain them by addressing the Kansas City Study of Adult Life, 716 Railway Exchange Building, Kansas City 6, Missouri.

[5]For reasons related to the content of the interview schedules and the analysis of data, it was decided to reverse the sequence of interviews with pick-ups. After Wave VI any respondent who had refused to be interviewed any time during the research was approached with the interview for Wave VI. If this interview was completed, he was approached at still a later date with the interview for Wave V, and so on. If a "pick-up" once again became a refusal on any interview, he was not approached during Wave VII. Therefore, all pick-ups listed for Wave VII have completed the whole series of interviews.

[6]Some also sent cards in return, but one wife felt miffed when one of her husband's cards was not addressed to her too.

[7]The statistical analysis of data was done by a two-way analysis of variance by ranks using the F test for significance. The formula used to obtain the F score is: $\text{Total} = \frac{m(n^3-n)}{12}$

$$\text{Between} = \frac{\sum(\text{Col}_1)^2 + \ldots (\text{Col}_{1+n})^2}{m} - \frac{mn(n+1)^2}{4}$$

Within = Total - Between

Using the results of the above formula where m is the number of rows and n is the number of columns then:

$$\underline{F} = \frac{\text{Sum of squares between columns}/n\text{-}1}{\text{Sum of squares within columns}/(n\text{-}1)(m\text{-}1)}$$

In the event of a mxn contingency table in which each row entry in a column receives the same rank as other rows in the column, the within group sum of squares will be equal to 0--that is, $\underline{F}$ is of maximum size. This relationship is evidenced by simple inspection of the table under consideration.

INTELLECTUAL CHANGE IN THE SENIUM

Robert W. Kleemeier, Ph.D.
Washington University, St. Louis

Until recent years it has been generally assumed that intellectual development and decline ran a relatively smooth and predictable course. Childhood was a period of rapid, uniform intellectual growth followed by a period of slower development during adolescence and perhaps early adulthood. After reaching its peak in the late teens or early twenties a gradual decline appeared. This intellectual decline was thought of as continuing at a constant rate throughout adult life until death intervened.

The factual basis for this conceptualization was derived mainly from the many cross-sectional studies of intellectual development using measures obtained from standardized tests of intelligence. The great majority of these studies were performed on children, considerably fewer on adults, and practically none on very old or senescent individuals. Such cross-sectional studies, while useful for normative purposes, only reveal performance differences between age groups. Nevertheless, they have been commonly interpreted so as to imply that these data are actually valid indicators of age changes. The fact remains, however, that age change can only be measured by longitudinal studies, i.e., the repeated testing of the same individuals at successive intervals during the life span. Fortunately some longitudinal data are available. Most important of these are the Bayley and Oden (1955) follow-up study of gifted children now in the mid-years of life, and Owens' (1953, 1962) retesting of college students thirty and forty years after their original examination. Neither of these studies report early decline in performance on tests of intelligence; indeed, they indicate continued growth rather than decline in the years of middle adulthood. These findings suggest that the drop in scores commonly reported on cross-sectional studies may reflect the environmental advantages of younger cohorts rather than the intellectual decline of their seniors.

In view of these results, the question remains at what point does the developmental curve turn downward, if indeed it does. It is not quite so absurd as it seems to question whether or not intellectual deterioration occurs in the normal course of life. Bear in mind that the two studies mentioned above do not support the hypothesis of intellectual decline in mid-adult life. Furthermore, a preliminary analysis of Moosehaven data, which will be reported on in more detail later, showed no significant change in performance in a group of men, aged 65 to 92, after a mean inter-test interval of 2.2 years (Arms, 1953). Indeed, the 26 younger subjects in this study, aged 65 to 75, showed a slight, but statistically insignificant improvement in performance on the second test. Surely these results would not allow one to predict confidently that deterioration would be found over a longer time interval.

Decline in Performance

The above analysis was a part of a study which was started at Moosehaven, a fraternal home for the aged, in 1949 and has been carried on there ever since. In fact the latest round of testing was conducted this summer on survivors of this study group (Arms, 1953; Kleemeier, Justiss, Jones and Rich, 1961; Obrist, Henry, and Justiss, 1961; Obrist, Busse, Eisdorfer, and Kleemeier, 1962). In all approximately 200 male residents of this institution were tested one or more times on the Wechsler-Bellevue Intelligence Scale for Adults. All were over age 65 and ranged up into the 90's.

In the 12 year period for which data are now available, 70 subjects were tested two or more times, with intervals of several years between tests. Thirteen of these were each tested four times. Since these latter subjects had received the most tests, their results were chosen for preliminary analysis, and are presented in Figure 1.

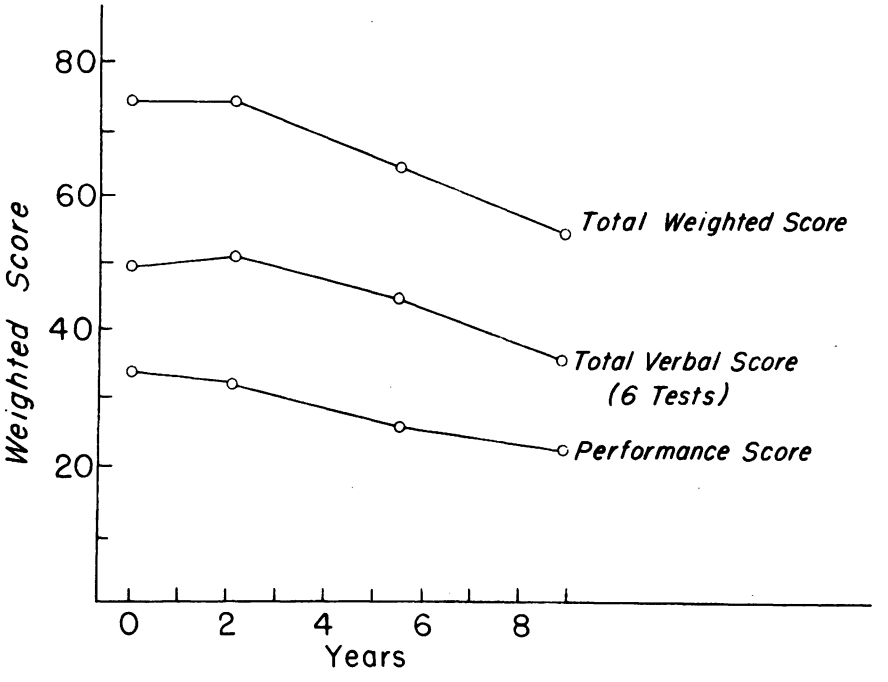


Fig. 1. Mean performance of 13 aged men on the Wechsler-Bellevue Intelligence Scale, Form I, on four successive tests.

This figure shows a decline in performance appearing by the third test; although, as would have been expected from the earlier analysis (Arms, 1953), no such trend would have been anticipated from the first two tests alone. Analyses of variance of these results show that the score differences on all three measures shown are significant at the .001 level.

These curves pose a problem. Why should there be a change in the rate of decline at the end of the second test? This is particularly curious, because of the wide age range of subjects represented in the curves. Does it mean that all subjects, regardless of age, experience a change at the end of the second test; or are some, but not all, subjects influencing the rate of change?

Individual curves

An answer is suggested in Figure 2, in which are shown the curves of nine subjects who were included in Figure 1 and who were living on January 1, 1961. Here we see that while there is a general drop in test score, considerable individual variation exists. Furthermore, the shape of the curve seems to be independent of the age of the subject.

Compare these curves with those shown in Figure 3, which represent the performance of four subjects who died shortly after the final test of the series. Each of these is characterized by a marked drop in performance and is distinctly different from those shown in Figure 2.

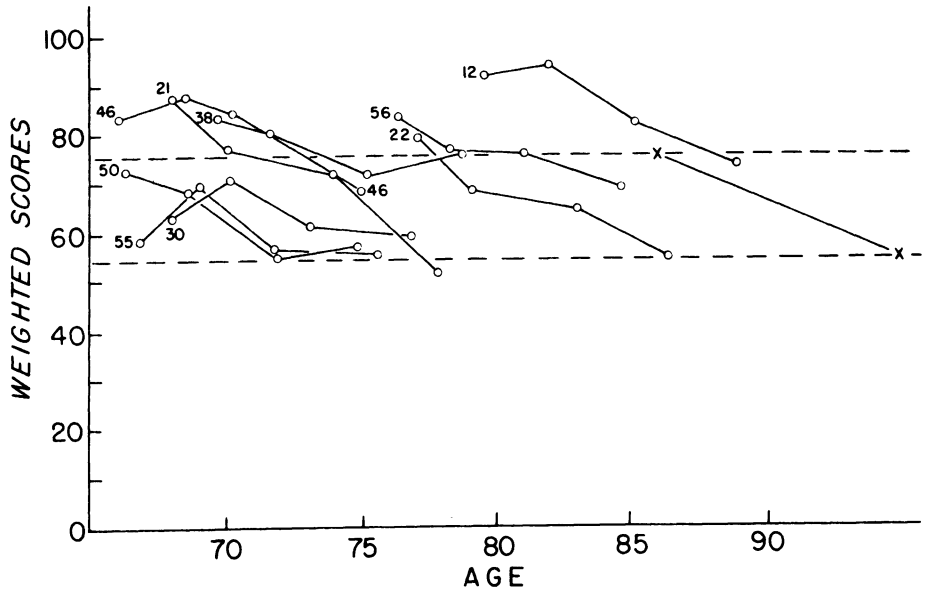


Fig. 2. Successive total weighted Wechsler-Bellevue scores, by age, of nine aged male subjects who were living on January 1, 1961. Upper dashed line represents mean first test score of all 13 subjects who were given four tests; lower line represents mean score of last test. The diagonal connecting these two represents the average rate of decline over the mean interval (8.99 years) between the first and last test.

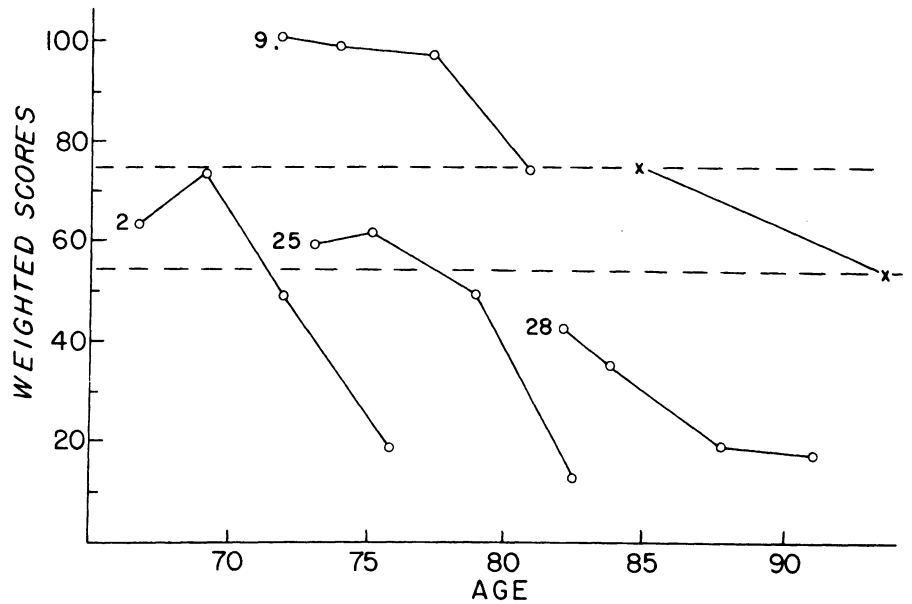


Fig. 3. Successive total weighted Wechsler-Bellevue scores, by age, of four male subjects who died before January 1, 1961. See legend Fig. 2 for details.

Terminal Drop

These curves suggest that factors related to the death of the individual cause a decline in intellectual performance, and that the onset of this decline may be detected in some instances several years prior to the death of the person.

In order to test this hypothesis, an analysis was made of data obtained from the 70 subjects who had been given at least two tests. Of these, 37 were living on January 1, 1961, and 33 had died prior to this date. Fortunately these two groups were of equal age at the time of their last test (deceased, 79.88 years; living, 79.56). Furthermore, there appeared to be no difference between the two groups in initial level of performance, thus any differential rate of decline between the groups could reasonably be attributed to factors related to the death of the one group.

Since the periods of time between successive tests were different for each subject, the mean annual rate of drop in total weighted score, in performance score and in verbal score were used to compare the two groups. Of the three only the performance score showed a significant difference (.001 level), and this was in the expected direction. Thus the mean annual drop in performance score was 1.99 points for the deceased group as opposed to .98 for the survivors. Since some of the survivors were also approaching death, and indeed some died within a few months of the arbitrary date chosen for the analysis, the reported differences between the two groups are conservative estimates of the possible influence of the "death factor" upon performance.

Institutionalization cannot be invoked as a cause of the obtained score difference, because the deceased group experienced its drop in performance over a shorter period of time than did the surviving group (4.58 years vs. 5.69 years; p less than .05).

It therefore appears that we have here strong evidence for the existence of a factor, which might be called terminal drop or decline, which adversely affects intellectual performance and is related to impending death of the aged person.

Other Evidence in Support of the Terminal Drop Hypothesis

Evidence in support of this hypothesis may be deduced from other studies, although they were not designed with this in mind. Most pertinent are the findings of Jarvik, Kallmann, Falek and Klaber (1957) in their study of changing intellectual functions of senescent twins. These authors compare curves of intellectual "change" during the senium derived from their own longitudinal study with comparable cross-sectional data. While both sets of curves show a progressive decline through the seventh and eighth decades, the rate of decline in the cross-sectional material is appreciably greater. Since the longitudinal data were based upon two tests given to each subject with an intertest interval of about eight years, it is obvious that at the time of the initial test subjects in this group had a minimum life expectancy of at least eight years. It is unlikely, therefore, that the initial test scores used in the calculation of the longitudinal curves were influenced by terminal drop. There is no such assurance for the subjects who provided the cross-sectional data. Indeed we would anticipate that the more aged these subjects were the greater the probability that some of them would already be closely approaching death and thus more likely to experience terminal decline in test performance. This could explain the increasingly poorer performance of the older age groups of cross-sectional subjects when compared to their age peers in the longitudinal group.

Additional evidence from the senescent twin study parallels our own findings. Original scores of subjects who survived through the taking of the second test were compared with the scores of those who died before the retest was administered. Although the results were statistically insignificant, the obtained mean score of the survivors was higher than that of the deceased group. It will be recalled that on a similar comparison we were able to demonstrate on the Moosehaven sample a greater mean annual rate of decline on the Wechsler-Bellevue total performance test score for the deceased group. We could not, however, show a difference in initial test performance between the two groups.

Again Jarvik and her associates noted that in 10 of twelve comparisons there was an increase in the mean intrapair test score differences on the 1955 tests over those obtained on the original (1947) tests obtained eight years prior to this time. Although again not significant statistically, the findings suggest that twins may become less alike in intellectual performance as they age in the senium. This is the kind of change which would be anticipated, if one of the twins in a pair began to experience terminal decline in anticipation of death, while the other remained in relatively good health.

Terminal Drop and EEG

The fact that decline in intelligence test score is related to the death of the person suggests the presence of an

organic factor related in some way to the health of the individual. An investigation by Obrist, Henry and Justiss (1961), based on the Moosehaven sample, sheds some light on this question. Their longitudinal study of 114 aged males shows a progressive slowing of the occipital alpha rhythm with increasing age. Furthermore, 28 cases who died shortly after their last EEG experienced a significant mean change of 0.6 cps. over a five year period, in contrast to a drop of only 0.3 cps. experienced over a seven year period by 28 surviving subjects. Thus the terminal drop found in intellectual performance seems to have its counterpart in EEG.

That intelligence test performance and EEG are indeed related was also demonstrated by Obrist and his associates (Obrist, Busse, Eisdorfer and Kleemeier, 1962). These investigators found a small but statistically significant correlation between intelligence test scores and alpha frequency on a sample of 115 Moosehaven subjects whose mean age was 77.0 years. In addition subjects whose tracings were characterized by diffuse slow activity exhibited significantly poorer intelligence test performance than those subjects whose tracings were free of this abnormality.

Although age could not be excluded as a source of at least a part of the common variance of EEG and intelligence, it was felt that health was a more significant contributor. This contention was supported by a failure to find any relationship between EEG and intelligence in aged persons living independently in the community. The fact that 26% of the Moosehaven group could be considered as convalescent because of some acute or chronic illness suggested that health rather than age per se might be responsible for the EEG-intelligence relationship, particularly since only a negligible number of the independently living older sample could be considered convalescent. Additional support for this point of view was provided by the finding that subjects showing some clinical evidence of arteriosclerosis exhibited a tendency both toward lowered intelligence test performance and mean alpha frequency when compared to a matching group without arteriosclerosis. Obrist and his associates suggest that "...variations in cardiovascular pathology more than any other factor, are responsible for the results obtained here, as well as those reported in the literature."

Conclusions

The significance of these studies can be summarized as follows. Great individual differences in intelligence test performance and in the amount of decremental change in this performance are found as age increases in the senium. These differences may be accounted for to a substantial degree by the presence of terminal drop or decline in performance associated with the death of the subject, but are not directly related to age in the senium. The presence of this death factor tends to associate intellectual decline to adverse change in health or organic integrity of the person rather than to so-called "normal" age change. It suggests further that any decline in performance not accounted for by terminal drop cannot simply be attributed to passage of time or to age alone, but must be explained in more explicit terms.

This reasoning leads to the defensible position that there is no evidence for age change in intelligence in the senium, and that changes which are found are better related to the physical state or health of the organism. Supporting this reasoning is the demonstrated fact of terminal drop both in intellectual performance and in EEG as well as the positive relationship found between EEG and intelligence test scores. The further linkage of both of these indices of decline to the presence of cardiovascular disease gives us all the more reason to believe that it is incorrect to refer to observed decrement in intelligence in the later years of life as the product of normal age change.

On the positive side these results permit the speculation that, in the absence of cardiovascular and other pathological conditions, post-maturational decline in intelligence does not occur. An additional point suggested by the data is that when intellectual decline does occur it is likely to follow an individual pattern, marked by relatively sharp decline, rather than a gradual wasting of ability as would be inferred from age curves of intellectual change based upon cross-sectional data.

REFERENCES

Arms, R. W. Wechsler-Bellevue retest after two years. In R. W. Kleemeier, Annual Report of the Moosehaven Research Laboratory. Orange Park, Fla., 1953, 69-73 (mimeo.).

Bayley, Nancy and Oden, Melita H. The maintenance of intellectual ability in gifted adults. J. Gerontol., 10, 91-107, 1955.

Jarvik, L. F., Kallmann, F. J., Falek, A., Klaber, M. M. Changing intellectual functions in senescent twins. Acta genet. 7, 421-430, 1957.

Kleemeier, R. W., Justiss, W. A., Jones, A. W. and Rich, T.A. Senescent changes in mental abilities. 1962 (mimeo).

Obrist, W. D., Henry, C. E., Justiss, W. A. Longitudinal study of EEG in old age. Excerpta Medica, International Congress Series No. 37, abs. 168, 1961.

Obrist, W. D., Busse, E. W., Eisdorfer, C. and Kleemeier, R. W. Relation of the electroencephalogram to intellectual function in senescence. J. Gerontol., 17, 1962, 197-206.

Owens, W. A.,Jr. Age and mental abilities: a longitudinal study. Genet. psychol. Monog., 48, 3-54, 1953.

Owens, W. A.,Jr. Personal communication. 1962.

DISCUSSION

Samuel W. Greenhouse, National Institute of Mental Health

I would first like to express my appreciation to the three speakers for having their papers ready at least three weeks prior to this meeting. This is quite an unusual occurrence.

Dr. Morrison and I have been the statist - cians involved in a multidisciplinary study of aging at the National Institute of Mental Health for the last five or six years, I am therefore familar with the problems of design and execution that have been raised in these papers, except possibly for the paper by Dr. Wheeler on maintaining a panel. For the most part these problems in methodology are not simple nor easy to solve.

At the outset, I must point out that I am sympathetic to the compromises usually made in solving these problems despite the fact that my comments may sound critical. This presentation is before a statistical group and it is necessary to keep in mind the normative procedure no matter how rigorous or impractical it may be to the investigator.

The major issue raised by Dr. Maddox is that of the selection of subjects. Anyone desiring to study characteristics of normal, healthy individuals aged 60 and over or the investigator interested in characterizing the process of normal aging, as we were in the NIH study, is immediately faced with the issue of how and where to obtain the subjects. Everyone of course is aware of what is necessary, namely, a representative sample selected at random from a given population. However, the lack of a frame or population listing of aged individuals living in the community would necessitate some more elaborate sampling procedure. Then the primary sample would have to be physically examined in order to obtain the sample of healthy i.e., free from disease or having asymptomatic, sub-clinical illness. Although many investigators feel reluctant about undertaking such a laborious and costly sampling scheme, it should be pointed out that it is possible and feasible. The alternative, of course, is getting the subjects the easiest way possible based on a publicity campaign. This usually means volunteers.

The most important consequence of not having a random sample is the uncertainty of the inferences that can be made from the study. Dr. Maddox discusses this very thoroughly. He refers to two inferential aims of inference: estimation of a characteristic or distribution of a variable in the population, and determining whether a relationship exists among a set of variables. He points out that generalizations of the first type require a sampling of the universe. He implies that generalizations of the second type "require, as a minimum, control of variables presumed to be relevant". I interpret this as meaning that the generalization of an observed relationship between variables in the sample can be made to the population of all normal, healthy aged despite the lack of a random sample, given that the condition of control on relevant variables is met. We have also assumed this to hold with respect to relationships among certain biological and physiological variables in the NIMH study. However, I should like to point out that this rests upon an implicit assumption; namely, that a correlation between two variables is unaffected by the manner in which subjects have been selected. For example, it would seem reasonable to assume that the existence and magnitude of the association between oxygen consumption in the brain and cholesterol level, say, would be little influenced by psychological and other factors that lead people to volunteer. On the other hand, it is less reasonable to suppose an association involving, say, blood pressure would be unaffected and, obviously, dangerous to assume relationships between psychological, psychiatric and social traits are unbiased in selected samples particularly volunteers. In short, all inferences either to the population of all normally healthy aged or to the aging process must be guarded and marked with caution.

My next comment is not peculiar to the Duke study but to most studies on aging which involve follow-up for a 5 or 10 year period. Why do most of these studies select subjects 60 or 65 and over? Why do they not start with age 50 or 55 and over? It has occurred to me if the object is to obtain information on the aging process, there may be more information on aging between ages 50 and 60 or 55 and 65 than between 65 and 75. I am aware, of course, of the additional difficulties in obtaining such subjects. The person 50 or 55 is still employed, and there exist many more obstacles in bringing him in for observation, for psychological testing and for physical examination. Despite this, it is possible, as demonstrated by the Kansas City study discussed by Dr. Wheeler.

I had not intended to make any comments on Dr. Wheeler's paper, primarily because I do not have much to say on the panel question. Very probably, this will be discussed fully by Dr. Taves. However, while Dr. Wheeler was speaking I was leafing through his manuscript again and came upon the footnotes which I had completely overlooked previously. In the footnotes, there is a reference to a test procedure used in the analysis of data which I believe is incorrect. The analysis is a two way analysis of variance of ranks in which I gather n columns are ranked from 1 to n in each of m rows. If we denote the "between column sum of squares" by B and the "within column sum of squares" by E, the author formed the ratio $\frac{B}{n-1} \Big/ \frac{E}{(m-1)(n-1)}$ and enters the tables of the F distribution with (n-1) and (m-1) (n-1) degrees of freedom. Now I believe that the denominator sum of squares is not distributed like Chi-square and is not independent of the numerator sum of squares. Therefore the ratio should not be compared with the F-distribution. The usual way in which the test is made is, for

large n and m, to compare the column sum of squares with the Chi-square-distribution having n-1 degrees of freedom. For small n (and/or m), there do exist tables with the exact .05 and .01 points, (see Siegel's Non-parametric Statistics.)

[Note added after the discussion. I should like to clarify the above remark since it can be interpreted as implying that a test statistic based on the ratio of the between sum of squares to the within sum of squares cannot or should not be used. No such implication was intended. The important point was that the above ratio should not be tested in the F-distribution with (n-1) and (m-1) (n-1) degrees of freedom. Such a test will tend to give a lower probability than the exact probability and thus lead to an actual significance level above the nominal one of .05 or .01. Actually, the Chi-square test appears to have a defect in the opposite direction. It tends to give a higher probability than the permutational probability and is therefore conservative, operating at a significance level below the nominal one.

Kendall (The Advanced Theory of Statistics, Vol. 1, First Edition, pp. 410-421) presents a test which he claims to be more accurate than the Chi-square test. This test is based on his coefficient of concordance, W. The ratio $(m-1)\ W/(1-W)$ is exactly equal to the ratio of the between mean square to the within mean square. However, the degrees of freedom with which one enters the F-table is reduced to

$$(n-1-\frac{2}{m}) \quad , \quad (m-1)\ (n-1) - \frac{2(m-1)}{m}$$

Even this test gives lower probabilities than the actual probabilities. Kendall suggests a continuity correction be applied to W. His recommended test is that the ratio $(m-1)\ W_c/(1-W_c)$ be entered into the F-distribution with the above adjusted degrees of freedom.]

Dr. Kleemeier's paper is somewhat different from the other two papers in that it presents a substantive result. I have two questions. Dr. Kleemeier started with 200 males, aged 65 and over, tested one or more times. Yet the data in support of the finding of intellectual deterioration before death is based on only 13 of these subjects. In order to protect himself against the argument of a special selectivity factor or other biased factor operating, he should present test results even if it is only 2 or 3 tests per subject and even if the subjects are still living, for all or a portion of the remaining 187 subjects.

Secondly I ask why in these 13 cases does deterioration come in every case with the third test. What is so unusual about test 3? Why didn't deterioration sometimes come at the second or fourth tests? The 13 individuals were first tested at different ages and showed varying durations of life after the first test. One could argue that it is taking the test a third time that produces the decline. Somehow there is needed a control testing procedure in which some individuals receive three tests in 6-8 years, others 2, others 4, others 5 and so on.

Again, as in my opening comment, I should like to thank the speakers for three excellent papers and for making them available long before this meeting.

THE FATE OF A PANEL OVER A FIVE-YEAR PERIOD - A CRITICAL REVIEW

Marvin Taves, Ph.D.*

The attempt by Wheeler** to analyze the sources of attrition in two samples of older adults represents an all too rare attention to methodological detail in social science research. The information presented can be useful in planning future longitudinal interview studies with older adults.

As Wheeler has pointed out, it is difficult to determine the limits of generalization from his first sample to other populations. In generalizing these findings it is important to note that the main sample is restricted to the middle socio-economic range of white noninstitutionalized relatively healthy urban persons. The second sample (referred to in the report as a quasi-sample) consists of persons aged 70 through 90 selected by interviewers from among neighbors, acquaintances, and "other sources", obviously not necessarily a random sample. Therefore, a more detailed description of the distribution of these samples and the characteristics of dropouts on such variables as education, occupation, and financial condition could help the reader determine the applicability of these findings in other situations which require the anticipation of sample attrition in the designing of a study.

Wheeler and his associates are to be commended for attempting to identify with some precision the various sources of attrition. Thus they held constant social class while examining dropouts by age and by sex.

Unfortunately the number of cases available often practically precluded statistically significant differences between categories under examination. Thus on page 6 one reads, "Out of 174 persons completing Wave I, eight--four males and four females--were lost through moves. The females, who moved were, however, concentrated in the lower-middle class, while males who moved were from both the upper-middle and the upper-lower, but not the lower-middle, classes." This conclusion is based on the data presented in Tables Ia and Ib which show that among the men 2 of the 4 dropouts by moves were in the upper-middle and 2 in the upper-lower social class, while among women all 4 dropped because of moves were in the lower-middle class.

Larger samples would most likely have reflected considerably different distributions in the contribution to dropping out from age, sex and social class. This conclusion is supported by the observed consistency in direction of differences on the independent variables.

Both the analysis and the report might have been strengthened through use of the directional hypothesis and one tailed rather than two tailed tests of significance. From the report I was not able to determine that this more powerful approach had been utilized.

I should like to encourage the investigator to give further attention to the quality of the interviewers. My own computation of the data presented in the report indicates that of the dropouts, 83 percent of the women and 58 percent of the men consisted of refusals. Death accounted for only 6 percent of the women and 28 percent of the men dropping out, while moves accounted for 11 and 14 percent respectively. This, in combination with the statement, "nearly all of the previously mentioned 'pick ups' were moved from the refusal category to the completion category through the efforts of this interviewer (i.e., an especially competent highly trained interviewer)".

I should like to congratulate the investigator for his careful analysis and clear report. At the same time I would suggest that it might be fruitful to investigate further, differences in dropouts between the major sample and the quasi-sample report.

Dr. Kleemeier* addressed himself to a significant substantive question, namely: is decline in intellectual performance during the senium caused by "factors related to death" rather than by age per se? He concludes, quite confidently, that, contrary to traditional expectations, "the decrement in intellectual performance may be accounted for to a substantial degree by . . . terminal drop" leading to death, rather than by normal age change. Therefore, a pertinent question for the reviewer is: Does available data warrant this degree of confidence, or even the conclusion? To answer, it is well to examine the quality of the data and its treatment.

The paper first points out the origin of its hypothesis. An examination of three studies of the mid-fifties suggested "continued (intellectual) growth rather than decline in the years of middle adulthood". This prompted the question of "whether or not intellectual deterioration occurs" at all normally, which generated the hypothesis investigated.

Kleemeier is to be commended for adding to his own carefully analyzed data those from other published reports. With the latter, it is necessary to recognize the need for caution in using data gathered for another purpose to test a given hypothesis.

To what extent do these studies confirm the terminal drop hypothesis? In my opinion, the studies cited lend support to the hypothesis, but not as firm support as might be thought from a reading of the paper reviewed. For instance, the 1957 study by Jarvis and others, listed as "most pertinent" in Kleemeier's paper, provides support only via secondary inference. That is, Kleemeier

*Chief, Division of Field Services, Special Staff on Aging, U.S. Department of Health, Education, and Welfare.

**Dr. Wayne Wheeler, University of Chicago Committee on Human Development, and Director of the Kansas City Study of Adult Life.

*Dr. Robert Kleemeier, Professor, Psychology Department, Program on Gerontology, Washington University, St. Louis, Missouri.

interprets the poorer performance of the cross-sectional sample than that of the longitudinal as reflecting terminal drop deterioration. He assumes that a number, sufficient to lower test results in the cross-sectional sample, were approaching death. This may have been true, and his reasoning may well be sound; nevertheless, the conclusion rests on secondary inference. Furthermore, both the longitudinal and cross-sectional data showed some progressive decline through the seventh and eighth decades. The cited directly supportive finding of the Jarvis et al study, that original scores of survivors exceeded those who died before the retest, is statistically nonsignificant. The increase in intra-pair differences between twins, although fairly consistent, is also statistically nonsignificant.

Turn next to the evidence cited from drop in EEG. The argument as presented is that, since EEG and intelligence were associated for the Moosehaven group where 26 percent were convalescent, but not for an independently living sample, there is evidence to suggest an organic base for terminal decline. This, as well as the clinical data showing some association between arteriosclerosis, measured intelligence and mean alpha frequency, would best be used to guide the refinement of the terminal drop hypothesis for further study rather than as proof of its correctness.

In the reviewer's opinion, the available literature cited does allow the logic and premise posed in Kleemeier's paper, but it offers mainly tangential (though supportive) evidence for testing the hypothesis in question.

Now note the original data presented. Results on two groups of the Moosehaven Study subjects are reported. The first group consists of 13 men, who, for reasons not stated, had been tested four or more times during the preceding 12 years. What biasing selectivity, if any, may have operated, is not noted. The four who died during the interval reportedly showed statistically greater decline on Wechsler-Bellevue scores than did the rest. The statistical test used is reported, but it is not clear whether the changes over all four tests, or only those between the first two and last two tests, were used in determining statistical significance.

These observations, that is, those data presented on the screen, quite correctly are presented not as a test but as generating a dual hypothesis: that (1) "factors related to death cause a decline in intellectual performance and (2) may be detected in some instances several years prior to . . . death." (No data on the latter.)

A second pair of groups consisting of 37 living and 33 deceased persons as of January 1, 1961, provided the data to test the above hypothesis. These 69 were apparently the total of those who had been tested at least twice. The differences in changes on the verbal performance and the scores of the Wechsler-Bellevue between these 37 survivors and the 33 deceased are the sole original data presented to test the hypothesis. The report states "of the three only the performance score showed a significant difference". After indicating effective control on age and institutionalization, it is concluded "that we have here strong evidence for the existence of a factor (terminal drop) . . . which adversely affects intellectual performance and is related to impending death of the aged person".

Since the concept, "terminal drop", is central to the entire paper, it might have been desirable to have the investigator's definition of the concept. How long a period does it involve? From statements on page seven it may be deduced that terminal drop covers less than eight years, while according to page five that it may be detected several years prior to death. Is it meant to identify a strictly functional deterioration or does the writer view the concept as having an organic and physiological referent?

To this reviewer the terminal drop hypothesis is hereby hardly verified. However, it remains an hypothesis worthy of further careful testing. Kleemeier has set forth the research model to be used. More adequate samples can be expected to confirm the terminal drop hypothesis, help determine at what point before death terminal drop may be detected, and to what extent and in what way it is associated with death and may help predict it.

If it is finally demonstrated that the traditionally claimed decline in intellectual performance during middle age and through senescence is incorrect, the practical implications are manifold. Retraining of displaced or redundant labor, or of those required to change jobs for reasons of health; periodic return to formal education by professionals and technicians, less emphasis on "new blood" in an organization, and reduced excuses for poor performance by healthy oldsters, are but a few examples.

* * *