STATISTICAL STANDARDS FOR ANALYZING DATA BASED ON COMPLEX SAMPLE SURVEYS

Monroe G. Sirken, B. Iris Shimizu, Dwight B. Brock, and Dwight K. French
National Center for Health Statistics

Introduction

We are preparing a manual of standardized procedures that will be used by mathematical statisticians in reviewing statistical reports that are prepared by analysts in the National Center for Health Statistics (NCHS). The rationale and procedures of the quality control program for reviewing the Center's published reports were described by Levy and Sirken [1972]. After the manual has been tested, it will be used to train the analysts. In this paper, we describe the standards and protocols for preparing the texts of statistical reports that are based on complex sample surveys and we summarize our preliminary efforts in applying these standards in the review process.

Vital and Health Statistics is one of the principal publications of the Center. The publication contains eleven series of reports, the combined series being often referred to as the "rainbow reports" since each series of reports has its own distinctively colored jackets. There are four series of reports which are based on data collected in complex sample surveys including the Health Interview Survey, the Health Examination Survey, the Institutional Population Surveys and the Hospital Discharge Survey. Some 20 to 25 substantive statistical reports are published annually in these four series.

Although the styles differ somewhat depending on the series and the analysts, basically, the format and composition of the four series of reports are quite similar. There are essentially three parts to a report. (1) Summary tables present the basic findings of the survey. (2) Appendices describe the statistical limitations of findings including the estimates of sampling error and, when available, estimates of measurement error. (3) The text, which is usually descriptive rather than analytical, summarizes and highlights the findings presented in the summary tables. Standards and protocols for each part of the report will be presented in the manual for reviewing statistical reports. We limit the scope of this paper, however, to a consideration of the statistical standards for reviewing the text which has presented, by far, the greatest difficulty and challenge.

We are well aware of the limitations of these initial efforts, and some qualifiers need to be made. In preparing these statistical standards, we limited our concern to developing statistical tests that would justify the kinds of statistical statements that are being made in the texts of statistical reports of the National Center for Health Statistics. Although we believe the standards might be applicable to descriptive reports prepared by other Federal statistical agencies, we cannot vouch for this possibility. These standards provide a basis for accepting or rejecting the statements that appear in the texts, but only from the viewpoint of sampling errors. We assume that estimates of sampling errors are available for all statistics presented in the summary tables, and we also assume that large sample normality assumptions apply. The standards do not at present consider the effects of nonsampling errors. Nor do these standards provide a means for determining whether statistical statements that are justified by the data and should have been made, in fact, are made in the texts of statistical reports.

Data based on complex sample surveys present difficult problems for analysts in the Center as well as in other Federal statistical agencies. Recently, Kruskal [1973] and Moore [1973] have commented on some of these common core problems. We are not satisfied with the tests we have developed, and hopefully better tests will evolve. In the meantime, we believe there is some virtue in establishing some standards if for no other reasons than to increase the comparability of reports prepared by different analysts and to improve communication between the analysts and the data consumers.

Statistical Standards

In reviewing the text of a report, the basic unit of analysis is the statistical statement. This is defined as a phrase, clause, sentence, or group of sentences which make inference(s) about population parameter(s) from statistics that are subject to sampling and/or measurement errors. A statistical statement may either contain an estimate of a parameter, or it may report the outcome of a test of a hypothesis.

The principal function of statistical statements is to describe and summarize the estimates that are presented in the summary tables of the report. Usually, the text will devote at least a paragraph to each summary table. Exhibit A is a model of the typical summary table. The stub of the table presents demographic variables, in this example, age and sex. The spread variable is usually a health variable, in this case hospital bedsize. Other examples of health variables include: type of acute conditions, cause of death, birthweight (of newborns), type of aid used by residents in nursing homes, etc. The cell entries in the table are estimates of morbidity rates for which the denominators are usually the sizes of the exposed-to-risk populations. Some examples of population morbidity rates are: percentages of the population with a specified health attribute, incidence and prevalence rates, etc. But the denominator is not necessarily the size of a population of persons. In the model table presented in Exhibit A, for example, the denominator of the average length of stay is the number of discharges.

EXHIBIT A. Average Length of Stay, by Sex and Age of Patient and Bed Size of Hospital: United States, 1967

Sex and Age	Hospital Bedsize					
	All Sizes	< 100 Beds	100-199 Beds	200-299 Beds	300-499	500 + Beds
Both Sexes	Average Length of Stay (in days)					
All Ages	8.4	7.5	7.6	8.7	8.9	10.3
< 15	5.5	4.1	4.8	5.2	5.9	8.4
15-44	6.2	5.1	5.5	6.4	6.7	7.7
45-64	10.1	8.5	9.2	10.2	10.7	13.1
65+	14.1	13.1	13.6	14.4	14.9	16.0
Male						
All Ages	9.0	7.5	7.8	9.3	9.7	11.4
< 15	5.5	4.1	4.8	4.9	6.0	8.2
15-44	7.3	5.4	6.1	7.9	8.3	9.8
45-64	10.2	8.2	9.0	10.3	11.3	13.1
65+	13.5	12.2	12.4	14.3	14.2	15.6
Female						
All Ages	8.1	7.5	7.5	8.2	8.3	9.5
< 15	5.5	4.2	4.7	5.5	5.8	8.5
15-44	5.8	5.0	5.3	5.9	6.1	6.9
45-64	10.1	8.8	9.3	10.1	10.2	13.0
65+	14.7	13.8	14.5	14.4	15.4	16.3

The application of statistical standards in reviewing the texts of statistical reports involves three distinct operations: (1) identifying statistical statements, (2) classifying statistical statements and (3) testing the validity of statistical statements.

The reviewer identifies statistical statements by bracketing sets of contiguous words in the text. Actually this involves two steps. First, the reviewer decides whether a set of words is a statistical statement because it makes an inference. Second, he identifies the words that begin and end every statement, such that none of the statements overlap.

After the reviewer identifies the statistical statements, he classifies them. According to our current classification scheme, which is based entirely on the Center's statistical reports, there are essentially five type of statistical statements. These types are listed and defined in Exhibit B. The illustrations of the types of statements presented in Exhibit B refer to morbidity rates that are displayed in Exhibit A.

Finally, the reviewer judges the validity of the testable statistical statements. Tests have been devised for each type of statement except for type 5. This type of statement is untestable either because it is ambiguous or it is not amenable to an existing test. Parsimony was a guiding principle in developing the typology of statements because we opted for a few general tests in preference to many specific tests. We were hopeful that type 5 statements would comprise a small portion of all statistical statements.

Simple statements do not involve testing hypotheses. Hence, the morbidity rates contained in these statements are subject only to reliability checks. We require that the coefficient of variation of the estimated rate be less than or equal to 25 percent. If fewer significant digits are given in the statistical statement than are shown in the summary table, we also require that the difference between the text figure and the estimate in the table either is less than one standard error of the estimate or less than five percent of the estimated value, whichever is smaller.

Single, multiple, and joint comparison statements make comparisons between two or more morbidity rates and hence involve tests of statistical hypotheses. The tests are made at the five percent level of significance and they are carried out as two-tailed tests except for those statements in which the analyst has specifically stated an interest in one-sided alternatives.

The test for a single comparison statement, which compares morbidity rates for a population containing two subdomains, is the usual test for significant differences between normal deviates. It is noteworthy that <u>not</u> all statements that compare morbidity rates for two subdomains are

EXHIBIT B. Definitions and Illustrations of Types of Statistical Statements

Type of Statement	Definition	Examples
1. Simple Statement	Infers the morbidity rate of a population.	1. The average length of stay for females was 8.1 days. 2. The average length of stay for persons under 15 in hospitals with 300-499 beds was about 6 days.
2. Single Comparison	Compares the morbidity rates for a population domain containing two subdomains.	1. The average length of stay for males was higher than for females. 2. Males had an average length of stay in the largest hospitals that was 1.9 days longer than the corresponding average for females.
3. Multiple Comparison	Compares morbidity rates for a population domain containing more than two subdomains.	1. The age group with longest average length of stay was the group 65 and over. 2. The average length of stay in the largest hospitals ranged from a low of 7.7 days in the 15-44 year age group to 16.0 days for those persons 65 and over. 3. Average length of stay for males increased for each successive age group. 4. Average length of stay tended to increase as hospital bedsize increased. 5. Average length of stay for males was higher in the largest hospitals than in the smallest.
4. Joint Comparison	Compares the morbidity rates between the subdomains for two or more population domains.	1. Males had a longer average length of stay than females in each of the three largest bedsize categories. 2. The average length of stay for males increased with each increase in age for all categories of hospital bedsize. 3. Average length of stay tended to increase within each age group as hospital bedsize increased.
5. Untestable Statement	An ambiguous statement or a statement for which a statistical test does not exist.	1. There exists a significant difference in the age distributions of average length of stay between males and females. 2. Comparable percentages among females are somewhat more variable across the age range than those for males.

classified as single comparison statements. Unless the analyst presents evidence in the text that he conjectured the hypothesis before viewing the data, these statements are classified as single comparison statements only if the two domains, such as male or female, comprise the entire population domain. Otherwise statements comparing the morbidity rates of two subdomains are classified as multiple comparison statements. For example, a statement comparing the average length of hospital stay for two bed size groups of hospitals is a multiple comparison statement because hospitals are classified into more than two bed size groups (see Exhibit A).

Multiple and joint comparison statements imply comparisons between more than two morbidity rates. With very few exceptions, which we will not describe here, statements of these types are tested by the Bonferroni method of multiple comparisons. This test is applied to the hypotheses implicated by the statistical statement. For instance, suppose that there are H possible pairwise comparisons between the morbidity rates comprising the population subdomains. Then the null hypothesis is that each of the H differences is equal to zero. The alternative hypothesis, however, could take a variety of forms: perhaps one, or two or even all of the H differences are nonzero. The alternate hypothesis, whichever form it takes, is determined by the statistical statement.

As indicated in Miller [1966], the Bonferroni method consists of a family of tests. Using the notation of Dayton and Schafer [1973],

let us define the probability of a nonzero family error rate by P(F); that is, P(F) is the simultaneous significance level for the defined family of tests. To avoid the additional assumption of independence of all component tests in the family, we use the Bonferroni inequality as an upper bound on the simultaneous significance level. Specifically, $P(F) \leq \sum_{i=1}^{H} \alpha_i$, where α_i is the significance level of the ith component test. This bound can be derived from Boole's inequality, a well-known result in probability theory--see, for example, Feller, [1968].

In our application of the Bonferroni technique to multiple comparisons, each component test is the usual test for significant differences between normal deviates, but with the α_i (as above) adjusted so that the simultaneous significance level is $\sum_{i=1}^{H} \alpha_i = .05$. For simplicity, this adjustment is made by setting each $\alpha_i = .05/H$, where, again, H is the total number of possible comparisons implicated by the statement.

Joint comparison statements may be viewed as combinations of single comparison statements and/or multiple comparison statements. Joint comparisons are tested by the Bonferroni method of multiple comparisons with $\alpha_i = \frac{.05}{\sum_{j=1}^{S} H_j}$, where H_j is the total number of possible comparisons among subdomains in the jth population domain, and S is the number of population domains covered by the statistical statement.

Experimental Test of Statistical Standards

It is one thing to devise a set of standards for statistical statements and another to apply these standards with a degree of reliability. Therefore, we designed and conducted an experiment in order to obtain a measure of the reliability of applying the standards. Another objective of the experiment was to estimate how often "untestable" and each of the other four types of statistical statements appear in the NCHS reports. A final goal will be to obtain preliminary estimates of the proportion of statistical statements in NCHS reports that are valid on the basis of these standards. However, the final part of the experiment has not yet been completed.

The experiment was based on eight recently published NCHS rainbow series reports. Two reports were selected from each of four series of reports that are based on data collected in complex sample surveys. A compact section containing about 50 to 70 statistical statements was randomly selected within each report, making a total of more than 400 statistical statements in the experiment. It should be noted that these eight reports were prepared by the analysts prior to the development of the standards that were being applied to them.

Six mathematical statisticians served as reviewers in the experiment. They represented a variety of statistical backgrounds ranging from a junior statistician to mid-level statisticians with Ph.D.'s in mathematical statistics. Each reviewer independently read the sample texts of the eight reports, identifying and classifying the statistical statements according to the protocols described earlier. The experiment, thus, is based on six independent reviews of the texts of eight reports. In addition, a seventh measurement was made. This seventh measurement, referred to as the "true" measurement in the following analysis, represents the majority opinion of the six reviewer observations. In those cases for which there was no majority opinion among the reviewers, the statements were referred for adjudication.

According to the "true" measurement, the text covered in the eight reports of the experiment contained 457 statistical statements. These statements are distributed by classification type in Table 1. About 10 percent of the statements are untestable, about 20 percent are simple statements and about two-thirds are comparative statements. The six reviewers identified 2684 statements, an average of 447.3 statements per reviewer. Table 1 also distributes these 2684 statements according to the types classified by the reviewers. The two distributions are in close agreement. Nevertheless, there were substantial differences among the reviewers as to the number of statements identified and as to classification of statements.

Table 1: Percent Distribution of Statement Types by "True" Classification and Reviewers' Classification of the Type of Statistical Statement

Type of Statistical Statement	"True" Classification	Reviewers' Classification
Number of Statements	457	2684
Total	100%	100%
1. Simple	22%	23%
2. Single Comparison	9%	9%
3. Multiple Comparison	31%	30%
4. Joint Comparison	26%	27%
5. Untestable	12%	11%

Let us note here that even though a reviewer may disagree with the "true" measurement on the classification of a statement, this does not necessarily imply that there would be disagreement on the validity of that statement. For example, while a reviewer may misclassify a

multiple comparative statement and call it a single comparative statement, it is possible that the difference implied by the comparison is not significant under either classification. However, since we have not yet tested the statements in the experiment, we make no further comments in this paper concerning agreements with respect to statement validity.

We turn our attention to the differences between the reviewers in identifying and classifying statements. The statements identified and classified by each reviewer were compared with the "true" measurements. The total of these comparisons, summed over all reviewers, are summarized in Table 2. From this table, we are justified in inferring differences among the reviewers, since, as we noted earlier, the "true" measurement represents the majority opinion of the six reviewers. Table 2 indicates that reviewers were subject to two kinds of errors; identification errors and classification errors.

Identification errors were committed by reviewers whenever statistical statements were entirely missed or when nonstatistical statements were erroneously identified as being statistical statements. Identification errors were also committed when reviewers merged two or more separate statistical statements into a single statement or divided a single statistical statement into two or more statements. In the former case, we counted two statistical statements as missed and one as erroneously identified, and in the latter case, we counted one statistical statement as missed and two statements as erroneously identified. If none of the reviewers had erroneously identified any statements, a total of six times the number of "true" statements, that is $6 \times 457 = 2742$ statements, would have been counted in the experiment. Henceforth, we shall refer to this number as the total actual number of statistical statements. According to Table 2, however, a total of 3206 statistical statements were counted in the experiment because the six reviewers identified only 2220 statements that were actual statistical statements, according to the "true" measurement. In addition, they erroneously identified 464 statements which were not statistical statements and failed to identify, or missed, 522 statistical statements. The difference between the missed and the erroneously identified statements represents a net identification error of -2.1 percent, or a net undercount of 2.1 percent of the total actual number of statistical statements. The sum of the missed and erroneously identified statements represents a gross identification error of about 36 percent. That is, the number of statements incorrectly identified by reviewers represents more than a third of the total actual number of statements contained in the experiment.

Classification errors were committed by reviewers when they misclassified the 2220 statements that they correctly identified. Thus, according to Table 2 the reviewers incorrectly classified about 17 percent of the 2220 statements which were correctly identified or about 12 percent of the 2742 total actual number of statistical statements in the experiment.

Next we consider the net and gross errors in identifying and classifying statements by type of statement. These errors are derived from Table 2 and they are presented in Table 3. For example, of the 612 statements that should have been identified and classified by reviewers as simple statements, 81 were missed entirely and 5 were identified but erroneously classified. On the other hand, the reviewers classified 87 nonstatistical statements as simple statements, and in addition, they misclassified 10 statistical statements as simple statements. Thus, the net identification error and the net classification error are $(87 - 81)/612 = 1.0$ percent and $(10 - 5)/612 = 0.8$ percent, respectively. Their algebraic sum, which we will refer to as the combined net error, is 1.8 percent. Similarly the gross identification and classification errors are $(81 + 87)/612 = 27.5$ percent and $(5 + 10)/612 = 2.5$ percent, respectively, and their sum, 29.9 percent, is the combined gross error. The errors presented in Table 3 for the remaining types of statements are calculated in a similar manner.

The absolute values of the combined net errors in Table 3 exceed 5 percent for two of the five statement types. The combined net error for the multiple comparative type is -5.3 percent, and for the untestable statements, it is -10.1 percent. With respect to the components of the combined net errors none of the absolute values of identification errors or classification errors is greater than 5 percent except for the net identification error for untestable statements, which is -12.6 percent.

The combined gross errors in Table 3 are large for every type of statement; they range from about 30 percent for simple statements to 100 percent for untestable statements. All of the gross identification and classification errors are about 25 to 50 percent, except for the gross classification error for simple statements, which is 2.5 percent.

For every type of statement, identification errors contributed more than classification errors to both the combined net errors and the combined gross errors. Entirely missing statistical statements and erroneously enumerating nonstatistical statements were relatively minor identification problems compared to the problem of setting the bounds for the statements. Merging two or more statistical statements into a single statement and dividing a single statistical statement into two or more statements were both rather common types of identification errors for testable statements. On the other hand, errors due to dividing statements were much more common than those due to merging statements in identifying untestable statements because the reviewers often identified each testable part of an untestable statement as a separate statistical statement.

In closing we feel obliged to note the ways in which limitations in the execution of the experiment may have contributed to the large gross errors that were detected. The manual was

in a draft form and it became necessary to make some changes in the protocols during the experimental period. Also, two of the six reviewers had virtually no experience with the standards prior to the experiment. Since we view the experiment as a pretest of the standards, we anticipate that if we were to repeat the experiment the measurement errors would be substantially smaller than those presented in this paper.

Table 2: Comparison of Reviewers' Classification with the "True" Classification of Statistical Statements.

"True" Classification of Type of Statistical Statement	Reviewers' Classification						
	Total	Statistical Statements Missed by Reviewers	Type of Statistical Statement				
			1 Simple	2 Single Comparison	3 Multiple Comparison	4 Joint Comparison	5 Untestable
Total	3206	522	623	253	807	715	286
Non-Statistical Statements Identified by Reviewers	464		87	42	129	136	70
1. Simple	612	81	526	2	1	1	1
2. Single Comparison	258	48	1	171	25	8	5
3. Multiple Comparison	852	168	5	14	583	60	22
4. Joint Comparison	702	115	2	15	58	465	47
5. Untestable	318	110	2	9	11	45	141

Table 3. Net and Gross Errors in Identifying and Classifying Statements by Type of Statistical Statement

Type of Statement	"True" Number of Statements	Percent Net Errors			Percent Gross Errors		
		Combined	Identification	Classification	Combined	Identification	Classification
1. Simple	612	1.8	1.0	0.8	29.9	27.5	2.5
2. Single Comparison	258	-1.9	-2.3	0.4	65.5	34.9	30.6
3. Multiple Comparison	852	-5.3	-4.6	-0.7	57.9	34.9	23.0
4. Joint Comparison	702	1.9	3.0	-1.1	69.4	35.8	33.6
5. Untestable	318	-10.1	-12.6	2.5	101.3	56.6	44.7

References

Miller, R. G., [1966]. Simultaneous Statistical Inference. McGraw-Hill Book Co., Inc., New York.

Feller, W., [1968]. An Introduction to Probability Theory and Its Applications, Volume I (3rd Editn.). John Wiley and Sons, Inc., New York.

Levy, P. S. and Sirken, M. G., [1972]. "Quality Control of Statistical Reports." Proc. Am. Statis. Assoc. Social Statis. Sec., 356-359.

Dayton, C. M. and Schafer, W. D., [1973]. "Extended Tables of t and Chi Square for Bonferroni Tests with Unequal Error Allocation." Journal of the American Statistical Association, 68, 78-83.

Kruskal, W., [1973]. "The Committee on National Statistics." Science, 180, 1256-1258.

Moore, G. M., [1973]. "On the 'Statistical Significance' of Changes in Employment and Unemployment." Statistical Reporter, 137-139.

SOME USEFUL STRATEGIES IN THE MULTIVARIATE ANALYSIS OF DATA FROM COMPLEX SURVEYS

Gary G. Koch, Daniel H. Freeman, and Jean L. Freeman, University of North Carolina

1. INTRODUCTION

For many large scale surveys like those conducted by the U.S. Bureau of the Census and the National Center for Health Statistics, data are obtained through complex designs often involving both clustering and stratification as well as multi-stage selection. Moreover, sub-population (or domain) characteristics are estimated by appropriate ratio methods. As a result, standard methods of multivariate analysis (which assume simple random samples) are not directly applicable. On the other hand, since sample sizes in such situations are usually very large, it generally can be assumed that the various estimates of sub-population characteristics do approximately have multivariate normal distributions with covariance matrices which can be consistently estimated by either direct or replication methods. Thus, a weighted least squares approach can be used to investigate various relationships among these estimates and test appropriate hypotheses. This paper is concerned with the application of this methodological strategy for analyses involving:

1. comparisons among cross-classified sub-populations,
2. evaluations of the existence and nature of trends.

2. METHODOLOGY

2.1. Definitions and Preliminaries.

Population: A set of N individuals indexed by the subscript $i = 1,2,\ldots,N$.

Sample Design: Probability random sample of size n with $\tau = 1$ trial for the measurements on each selected individual. As indicated by Cornfield [1], this sample design can be characterized by random variables U_i where

$$U_i = \begin{cases} 1 & \text{if population element } i \text{ is in sample} \\ 0 & \text{otherwise} \end{cases}.$$

In this context, $\phi_i = E\{U_i\}$ represents the probability of selection for the i-th population element and $\theta_{ii'} = E\{U_iU_{i'}\}$ represents the probability for the joint selection of both the i-th and i'-th population elements. Unless stated otherwise, we shall usually assume that sampling is without replacement in which case $\theta_{ii} = E\{U_i^2\} = E\{U_i\} = \phi_i$; however, the $\theta_{ii'}$ for $i \neq i'$ must be determined by appropriate calculations which correspond directly to the specific method of selection -- although for some complex designs this can potentially involve very difficult mathematical problems.

Measurement Process:

a. Self-enumeration. In other words, each individual in the sample responds individually and independently to the survey measurement process. Eg., mailed questionnaires or personal contact situations in which the role of the enumerator or interviewer is minimized.

b. Random assignment of interviewers. There is a fixed population of B interviewers which are available for assignment to the n sampled individuals. It will be assumed that this process is undertaken at random with n_h individuals being associated with the h-th interviewer. Thus, $n = \sum_{h=1}^{B} n_h$.

The assignment of interviewers can be characterized by random variables T_{hi}, where

$$T_{hi} = \begin{cases} 1 & \text{if interviewer h is assigned to individual i given i is in sample} \\ 0 & \text{otherwise,} \end{cases}$$

and their corresponding joint probability distribution.

Basic Observational Unit: The random variable $Y_{hit}^{(\xi)}$ represents the measurement corresponding to the ξ-th attribute (where $\xi = 1,2,\ldots,m$ indexes the m attributes) of the i-th individual in the population with respect to the h-th interviewer. The subscript t indexes a conceptual sequence of replications of this overall measurement process for any specific individual in the sample. Henceforth, the ξ-superscript will usually be dropped for notational convenience.

2.2. Model. In the spirit of the approach of Wilk and Kempthorne [11], one model of interest involves assuming that the Y_{hit} can be represented as an additive function of an overall mean, a fixed main effect due to the i-th individual, a fixed main effect due to the h-th interviewer, a fixed interaction effect due to the combination of the h-th interviewer and i-th individual, and a random residual effect corresponding to the combination of the h-th interviewer, i-th individual, and t-th trial. In this regard, we have the structure

$$Y_{hit} = \bar{Y} + B_h + H_i + (BH)_{hi} + Z_{hit}$$

where $\bar{Y}$, B_h, H_i, and $(BH)_{hi}$ are determined in the following manner from $Y_{hi} = E\{Y_{hit}\}$

$$\bar{Y} = \frac{1}{NB} \sum_{h=1}^{B} \sum_{i=1}^{N} Y_{hi}$$

$$B_h = \frac{1}{N} \sum_{i=1}^{N} (Y_{hi} - \bar{Y}) = (\bar{Y}_h - \bar{Y})$$

$$H_i = \frac{1}{B} \sum_{h=1}^{B} (Y_{hi} - \bar{Y}) \equiv (Y_i - \bar{Y})$$

$$(BH)_{hi} = (Y_{hi} - B_h - H_i - \bar{Y}) = (Y_{hi} - \bar{Y}_h - Y_i + \bar{Y})$$

$$Z_{hit} = (Y_{hit} - Y_{hi}).$$

With respect to measurement error models like those developed by the U.S. Bureau of the Census (see [4]), the $\{Z_{hit}\}$ reflect trial to trial variation in the context of all potentially observable measurements of a particular attribute

for any specific individual and interviewer combination. Thus, this source of variation represents intrinsic response error which is due to factors which are not under control with respect to the sampling selection and measurement data collection processes. Such response errors may be due to structural weaknesses or vagueness in the definition of the phenomena being measured for an individual (eg., attitudes related to political opinions, consumer taste preferences, etc.) or a consequence of some underlying stochastic process (occurrences of motor vehicle accidents and subsequent injuries, outcomes pertaining to judgments in court cases, survival subsequent to diagnosis and treatment for disease, pregnancy outcomes like birthweight, etc.). In view of these considerations, we shall henceforth assume that the $\{Z_{hit}\}$ are mutually uncorrelated and $\mathrm{Var}\{Z_{hit}\} = E\{Z_{hit}^2\} = \eta_{hi}^2$.

The $\{T_{hi}\}$ give rise to external response error with respect to the fact that there is controlled variability in the basic measurement process to the extent that different interviewers can potentially tend to report different observations for a particular attribute of a specific individual. The nature of this source of error is characterized by the $\{B_h\}$ and $\{(BH)_{hi}\}$. Since these quantities are rather difficult to manipulate in general terms, we shall henceforth assume that they are unimportant and can be neglected; i.e., we shall assume $B_h=0$ for all h and $(BH)_{hi}=0$ for all h,i. For other discussion, see Koch [7].

Finally, the $\{U_i\}$ reflect sampling error since their joint probability distribution characterizes the selection aspects of the survey design in the sense of which individuals are in the sample and which are not.

a. Special Case for Only Sampling Errors. In this situation, $\eta_{hi}^2 = 0$ for all h,i and thus $Z_{hit} \equiv 0$. As a result, the model becomes

$$Y_{hit} = \bar{Y} + H_i.$$

Examples might include determinations of whether a product was defective or not in certain types of acceptance (inspection) sampling or the cost of purchases of stock items in inventory samples.

b. Special Case for Only Response Errors. In this situation $H_{hi} = 0$ for all h,i. As a result, the model becomes

$$Y_{hit} = \bar{Y} + Z_{hit}.$$

Examples might include determinations of the distribution of the ratio of the harmonic vs. geometric mean for the observations corresponding to the faces of four twelve-sided dice, determinations based on repeated simulations of a given stochastic process which are all based on the same computer random number generator (although with different starting points), repeated experimental observations on different samples from basically the same bacteria culture, repeated observations on the preferences of a specific individual with respect to (blind) paired comparisons of particular food or beverage products. However, as will be argued later, perhaps the most important situations of this type involve single observations on distinct individuals who belong to a matched set based on twin relationships, other family relationships, or equivalence with respect to several demographic or other characteristics.

2.3. Other Assumptions.

a. No interaction in the measurement process in the sense that $E_t\{Y_{hit} | \text{any specification of } \{U_i\} \text{ and } \{T_{hi}\}\} = Y_{hi}$ and hence conditional expected value notation of the type $E_t\{\ |\ \}$ will not be used in any of the remaining discussion.

b. The measurement process is unbiased in the sense that the population mean $\bar{Y}$ (or the population total $Y = N\bar{Y}$) is the population parameter of interest.

2.4. Linear Sample Statistics. Let us consider the model described in (2.2) with respect to the statistics

$$y_t = \frac{1}{B} \sum_{i=1}^{N} \sum_{h=1}^{B} W_{hi} U_i T_{hi} Y_{hit}$$

which is a linear combination of the observed elements in the sample with the W_{hi} being known specified coefficients. On the basis of previous assumptions, we have

$$E\{y_t\} = \frac{1}{B} \sum_{i=1}^{N} \sum_{h=1}^{B} W_{hi} \phi_i \lambda_{hi} Y_{hi}$$

where $\phi_i = E\{U_i\}$ and $\lambda_{hi} = E\{T_{hi} | U_i = 1\}$. In the remainder of this paper, we shall assume that the weights are the reciprocals of the probabilities associated with each (h,i) combination; i.e., $W_{hi} = (1/\phi_i\lambda_{hi})$ so that y_t represents a generalized Horvitz and Thompson [5] estimator which accounts for non-uniform assignments of interviewer effects. We shall now assume that the $\{\lambda_{hi}\}$ are uniform in the sense that $E\{T_{hi} | U_i=1\} = (1/B)$ in which case we can write y_t in the usual Horvitz-Thompson form

$$y_t = \sum_{i=1}^{N} W_i U_i \{ \sum_{h=1}^{B} T_{hi} Y_{hit} \} = \sum_{i=1}^{N} \left(\frac{1}{\phi_i}\right) U_i Y_{it}$$

where $W_i = (1/\phi_i)$ and $Y_{it} = \{ \sum_{h=1}^{B} T_{hi} Y_{hit} \}$ so that

$$E\{y_t\} = \sum_{i=1}^{N} Y_i = Y = (N\bar{Y})$$

since $E_{t,T}\{Y_{it}\} = Y_i$ with $E_{t,T}\{\ \}$ being interpreted as expected value with respect to both conceptual repeated trials as well as interviewer assignments.

As indicated in Koch [7], the variance of y_t under the model (2.2) in conjunction with the assumptions (2.3) can be written as

$$\mathrm{Var}\{y_t\} = \frac{1}{n}\{(SRV) + (n-1)(CRV)\} + \{(SV)\}$$

with

$$(SRV) = \text{Simple Response Variance} = N^2[\frac{1}{N}\sum_{i=1}^{N} (\bar{\phi}/\phi_i)\ E_{t,T}\{(Y_{it}-Y_i)^2\}];$$

$$(CRV) = \text{Correlated Response Variance} = \frac{N^2}{N(N-1)} \sum_{i \neq i'}^{N} \psi_{ii'} E_{t,T}\{(Y_{it}-Y_i)(Y_{i't}-Y_{i'})\}$$

(SV) = Sampling Variance

$$= \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{\theta_{ii'}}{\phi_i \phi_{i'}} - 1\} Y_i Y_{i'} \text{ ; where}$$

$\bar{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi_i = (n/N)$, $\bar{\theta} = \frac{1}{N(N-1)} \sum_{i \neq i'}^{N} \theta_{ii'} = n(n-1)/N(N-1)$, and $\Psi_{ii'} = \frac{\theta_{ii'} \bar{\phi}^2}{\bar{\theta} \phi_i \phi_{i'}}$.

In accordance with the assumptions regarding intrinsic response error (due to uncontrolled observational factors) and external response error (due to interviewer effects, etc.) given in 2.2 and the assumptions in 2.3 it follows that there is no correlated response variance component since CRV = 0 under these conditions and $\text{Var}\{y_t\}$ simplifies to $\text{Var}\{y_t\} = \frac{1}{n}\{(\text{SRV})\} + \{(\text{SV})\}$ where $(\text{SRV}) = N^2 \frac{1}{N} \sum_{i=1}^{N} (\bar{\phi}/\phi_i) \eta_i^2$ with η_i^2 being defined by $\eta_i^2 = \frac{1}{B} \sum_{h=1}^{B} \eta_{hit}^2$.

2.5. Estimators for the Variance of a Linear Sample Statistic. Here, we shall restrict attention to the usual Horvitz-Thompson statistic $y_t = \sum_{i=1}^{N} (\frac{1}{\phi_i}) U_i Y_{it}$ under the assumptions given in (2.2)-(2.4) with respect to the uncorrelated nature of intrinsic response errors and the nonexistence of external response errors.

a. Direct Methods. A lower bound estimator for the variance of y_t is the Horvitz-Thompson quadratic statistic

$$(\hat{SV}) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{1}{\phi_i \phi_{i'}} - \frac{1}{\theta_{ii'}}\} U_i U_{i'} Y_{it} Y_{i't}$$

for which $E\{(\hat{SV})\} = \frac{N^2}{n}\{\frac{1}{N} \sum_{i=1}^{N} (\frac{\bar{\phi}}{\phi_i} - \frac{n}{N}) \eta_i^2\} + (\text{SV})$.

Similarly, an upper bound estimator for Var t_y is

$$(\hat{\hat{SV}}) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{1}{(1-\phi_i)(1-\phi_{i'})}\}^{1/2} \cdot \{\frac{1}{\phi_i \phi_{i'}} - \frac{1}{\theta_{ii'}}\} U_i U_{i'} Y_{it} Y_{i't}$$

for which

$$E\{(\hat{\hat{SV}})\} = \frac{1}{n}\{(\text{SRV})\} + \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{1}{(1-\phi_i)(1-\phi_{i'})}\}^{1/2} \cdot \{\frac{\theta_{ii'}}{\phi_i \phi_{i'}} - 1\} Y_i Y_{i'} .$$

Although $(\hat{SV})$ and $(\hat{\hat{SV}})$ appear to be quite different, it follows that if n and N are large and if n << N so that the terms $(1-\phi_i)$ which tend to behave like (1-n/N) can be replaced by 1's, then $(\hat{SV}) \approx (\hat{\hat{SV}})$. Finally, these considerations can be simplified if all the ϕ_i are equal to (n/N) in which case

$$E\{(\hat{SV})\} = \text{Var}\{y_t\} - (\frac{1}{N})(\text{SRV})$$

$$E\{(\hat{\hat{SV}})\} = \text{Var}\{y_t\} + \frac{n}{(N-n)}(\text{SV})$$

Moreover, if all the $\theta_{ii'}$ are equal to n(n-1)/N(N-1), then the expressions for $(\hat{SV})$ and $(\hat{\hat{SV}})$ also simplify to the familiar forms

$$(\hat{SV}) = N^2 (\frac{1}{n})(1-\frac{n}{N})\{\frac{1}{(N-1)} \sum_{i=1}^{N} U_i (Y_{it} - \bar{y}_t)^2\}$$

$$(\hat{\hat{SV}}) = N^2 (\frac{1}{n}) \frac{1}{(N-1)} \sum_{i=1}^{N} U_i (Y_{it} - \bar{y}_t)^2$$

where $\bar{y}_t = (y_t/N)$ is the estimator for the population mean (in this case, the ordinary sample mean) and $s^2 = \{\frac{1}{(N-1)} \sum_{i=1}^{N} U_i (Y_{it} - \bar{y}_t)^2\}$ is the sample variance estimator in the usual sense.

In summary, if sampling variance is the most important source of error, then $(\hat{SV})$ is the most appropriate estimator since $(\hat{\hat{SV}})$ is needlessly conservative. However, if intrinsic response variance is the most important source of error, then $(\hat{\hat{SV}})$ is the most appropriate estimator since $(\hat{SV})$ will tend to underestimate the actual variance in a potentially misleading manner.

b. Replication Methods. For many surveys involving complex multistage selection procedures, the numerical calculations associated with estimators like $(\hat{SV})$ or $(\hat{\hat{SV}})$ can require considerable effort with respect to programming as well as substantial computer time costs. The main reasons for this is that these expressions involve n^2 terms and the subscript i may be a vector subscript. Thus, in recent years, there has been considerable interest in the development of alternative estimation procedures for the variances of sample statistics. In particular, one such method which has been already used extensively by the National Center for Health Statistics as well as other institutions or organizations engaged in survey research is the method of balanced repeated replication (BRR) as discussed for example by Kish and Frankel [6], Koch and Lemeshow [8] and McCarthy [9]. The principal concept which governs the use of BRR is that variability of a statistic based on a total sample can be estimated in terms of the variability of that statistic across subsamples (called replications) which reproduce (except for size) the complex design of the entire sample. Hence, BRR has considerable appeal in those situations where clustering causes the underlying distribution theory for determining the $\theta_{ii'}$ as well as the computational effort for calculating $(\hat{SV})$ or $(\hat{\hat{SV}})$ to become impractical. One specific version of BRR is the method of balanced half samples. This procedure is characterized by a matrix H with elements h_{ik} defined by

$$h_{ik} = \begin{cases} 1 & \text{individual i is in k-th half sample} \\ 0 & \text{otherwise} \end{cases}.$$

For each half sample, we form the estimator $y_{tk} = 2\sum_{i=1}^{N}(\frac{1}{\phi_i})U_i h_{ik} Y_{it}$ which is directly analogous to $\bar{y}_t$ with respect to estimating the population total Y. The resulting estimator V for the variance of y_t is $V = \frac{1}{L}\sum_{k=1}^{L}(y_{tk} - y_t)^2$ where L denotes the number of half sample partitions. In this context, it should be noted that the appropriate choice of the matrix $\underset{\sim}{H}$ represents a very important feature of this method for determining the estimator V. Some efficient strategies for this purpose are described in [9]. Finally, it should be recognized that this method of estimating variance primarily pertains to those situations where there are no important sources of external response errors (eg., interviewer effects) as assumed in most parts of this paper. However, appropriate modifications with respect to the definition of $\underset{\sim}{H}$ are certainly within the scope of the general BRR approach for constructing replication estimators of variance which reasonably reflect this source of error as well as intrinsic response error and sampling error.

2.6. Estimators for the Covariance of Two Linear Sample Statistics. Suppose $y_t^{(\xi)}$ and $y_t^{(\xi')}$ correspond to the Horvitz-Thompson statistics for estimating the population totals corresponding to the ξ-th and ξ'-th attributes respectively. Then the methods described in (2.5) can be used to determine estimators for $\mathrm{Var}\{y_t^{(\xi)}\}$, $\mathrm{Var}\{y_t^{(\xi')}\}$, and $\mathrm{Var}\{y_t^{(\xi)} + y_t^{(\xi')}\}$ where

$$y_t^{(\xi)} + y_t^{(\xi')} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i(Y_{it}^{(\xi)} + Y_{it}^{(\xi')})$$

is the Horvitz-Thompson statistic for estimating the sum of the population totals associated with the ξ-th and ξ'-th attributes in terms of the respective sums for individuals who are selected in the sample. However, this means that an estimator for Cov $y_t^{(\xi)}, y_t^{(\xi')}$ can be directly obtained from the identity relationship

$$\mathrm{Cov}\{y_t^{(\xi)}, y_t^{(\xi')}\} = \frac{1}{2}[\mathrm{Var}\{y_t^{(\xi)} + y_t^{(\xi')}\} - \mathrm{Var}\{y_t^{(\xi)}\} - \mathrm{Var}\{y_t^{(\xi')}\}]$$

by replacing the respective variance expressions by their corresponding estimators. Thus, an (m x m) estimated covariance matrix V can be determined for the joint set of estimators for the population totals corresponding to m attributes by using a computer program which calculates estimates of variance on individual univariate variables in conjunction with a variable sum operation and the previously indicated identity.

2.7. Estimates for Domain Means and Other Ratio Statistics. The term domain refers to subclasses derived from a particular response variable which is measured during the survey (ie., a posteriori with respect to selection) and which takes on categorical values (either directly as with marital status or indirectly after grouping as with age). The term strata refers to subclasses derived from a factor variable which is presumed known for each individual in the population prior to the undertaking of the survey (eg., region of the country, or urban vs. rural, etc.). Moreover, in most applications, several strata-type variables are directly related to the nature of the selection process with corresponding effects induced on the joint distribution of the U_i in some manner (ie., separate independent random samples are usually obtained from each subpopulation corresponding to appropriate combinations of such strata-type variables).

Estimates of subpopulation totals for both domains as well as strata can be formulated in terms of indicator variables of the type

$$X_{hijt} = \begin{cases} 1 & \text{individual i is classified in j-th domain on t-th trial measured by h-th interviewer} \\ 0 & \text{otherwise} \end{cases}$$

where $j = 1,2,\ldots,s$ by forming statistics like

$$y_{jt} = \frac{1}{B}\sum_{i=1}^{N}\sum_{h=1}^{B} W_{hi}\phi_i\lambda_{hi}X_{hijt}Y_{hit}.$$

We shall assume that the domain classification process is affected neither by intrinsic response error nor external response error and is also unbiased in the sense of (2.3); ie., we have

$$X_{hijt} = X_{ij} = \begin{cases} 1 & \text{the individual i is always classified correctly in the j-th domain} \\ 0 & \text{otherwise.} \end{cases}$$

However, it should be recognized that the presence of such response errors is a definite possibility in many survey situations and can have potentially important effects on the statistical properties of estimators like y_{jt} as discussed in Koch [7]; eg., y_{jt} is not necessarily unbiased unless such assumptions apply. Finally, strata-type variables are viewed in this same framework by definition together with the fact that the actual values of the $\{X_{ij}\}$ are known a priori constants for each individual in the population.

As indicated in (2.4) we shall consider the Horvitz-Thompson type estimators

$y_{jt} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i G_{ijt}$ which can also be written in the form $y_{jt} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i G_{ijt}$ where $G_{ijt} = X_{ij}Y_{it}$.

However, in this context, the discussion in (2.5)-(2.6) can be applied to obtain an estimated covariance matrix V for the vector of domain total estimators $\underset{\sim}{y}'_t = (y_{1t}, y_{2t}, \ldots, y_{st})$. Similarly, these same considerations can also be applied to the multivariate case of m attribute variables by working with the composite vector

$$\underset{\sim}{y}'_t = [\underset{\sim}{y}_t^{(1)'}, \underset{\sim}{y}_t^{(2)'}, \ldots, \underset{\sim}{y}_t^{(m)'}].$$

In many situations, there is actually greater interest in domain means which are basically ratio estimates of the type

$$\bar{y}_{jt} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i G_{ijt} \Big/ \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i X_{ij} = (y_{jt}/x_{jt}).$$

If $\underset{\sim}{y}'_t = (y_{1t}, y_{2t}, \ldots, y_{st})$ and $\underset{\sim}{x}'_t = (x_{1t}, x_{2t}, \ldots, x_{st})$ then the set of domain means $\underset{\sim}{\bar{y}}_t = (\bar{y}_{1t}, \bar{y}_{2t}, \ldots \bar{y}_{st})$ can be written in the compound function framework outlined in Forthofer and Koch [2]

$$\underset{\sim}{\bar{y}}_t = \underset{\sim}{R}[\exp\{\underset{\sim}{K}(\log_e(\underset{\sim}{A}\frac{\underset{\sim}{y}_t}{\underset{\sim}{x}_t}))\}] \text{ where } \underset{\sim}{A} = \underset{\sim}{I}_{2s} \text{ (which is the (2s x 2s) identity matrix),}$$

$$K = \begin{bmatrix} 1 & 0 & \ldots & 0 & -1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & -1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & -1 \end{bmatrix}$$

and $R = I_s$ (which is the (s x s) identity matrix) and the $\log_e$ ( ) vector operation forms the vector of natural logarithms and the exp vector operation forms the vector of anti-logarithms.

An estimate of the covariance matrix for $\bar{y}_t$ which is based on the large sample Taylor series linearized approximation can be obtained with direct matrix multiplication operations as follows

$$\text{Var}\{y_t\} \hat{=} V_y = RD_q K D_a^{-1} AVA' D_a^{-1} K' D_q R'$$

where "$\hat{=}$" means "is estimated by" and where D_a represents the diagonal matrix with elements of the vector $a = A\begin{bmatrix} y_t \\ x_t \end{bmatrix}$ on the main diagonal and D_q represents the diagonal matrix with elements of the vector $q = \exp\{K[\log_e(a)]\}$ on the main diagonal. Estimated covariance matrices for other sets of compounded functions involving estimates of domain totals can be produced in an analogous manner; eg, differences between domain means, post-stratified means, certain types of vital rates based on life-table functions, and rank correlation type measures of association.

In summary, compound function operations can be used to compute the vector $\bar{y}_t$ of estimated means for any given set of domain subpopulations. The corresponding estimated covariance matrix V_y can be calculated by previously described matrix multiplication operations. Alternatively, an estimated covariance matrix V_y could also be obtained by using the replication methods described in (2.5b) although there are potentially certain problems with singularities with this approach for reasons which are outlined in [8]. Nevertheless, the estimator $\bar{y}_t$ and its estimated covariance matrix V_y can be obtained for any specific survey situation within the context of the general framework described in (2.1)-(2.7). The problem now is to consider a general approach for undertaking statistical inference with respect to $\bar{y}_t$.

2.8. Multivariate Analysis for Estimates from Complex Sample Survey Data. Let F denote a (g x 1) vector of statistics like the domain mean estimates $\bar{y}_t$ described in (2.7). Let V_F denote an appropriate valid and consistent estimate of the corresponding (g x g) covariance matrix for F obtained by methods like those described in (2.4)-(2.7).

The relationship between variation among the elements $F_1, F_2, \ldots, F_g$ of the vector F and certain aspects of the nature of various subpopulations (or domains and subdomains) can be investigated by fitting linear regression models to the vector F by the method of weighted least squares. This aspect of statistical analysis can be characterized by writing

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \ldots \\ F_g \end{bmatrix} \hat{=} \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1u} \\ x_{21} & x_{22} & \ldots & x_{2u} \\ \ldots & \ldots & \ldots & \ldots \\ x_{g1} & x_{g2} & \ldots & x_{gu} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_u \end{bmatrix} = X\,b$$

where X is the pre-specified design (or independent variable) matrix of known coefficients with full rank u, b is the (u x 1) vector of unknown parameters or effects, and "$\hat{=}$" means "is estimated by." This particular model implies the existence of a [(g - u) x g] matrix L which is orthogonal to X such that $\{f = L F = L X b \hat{=} 0\}$ represents a corresponding set of implied constraints. Thus, it follows that the covariance matrix of f can be estimated by $V_f = L V_F L'$. As a result, an appropriate test statistic for the goodness of fit of the model of interest is

$$Q = f' V_f^{-1} f = f'(L V_F L')^{-1} f$$

which is approximately distributed according to the χ^2-distribution with D.F.=(g - u) if the overall sample size n is sufficiently large that the elements of the vector F have an approximate multivariate normal distribution as a consequence of Central Limit Theory. Such test statistics are known as Wald [10] statistics and various aspects of their application to problems involving the multivariate analysis of multivariate categorical data are discussed in [2], [3]. Moreover, since the sample sizes associated with most complex sample surveys are generally very large so that it is reasonable to assume that the resulting estimates of population characteristics tend to have approximately normal distributions (as a consequence of Central Limit Theory), such Wald statistics provide a valid and potentially useful framework for the multivariate analysis of the resulting estimates. However, the actual manner in which this approach is undertaken involves a weighted least squares computational algorithm which is justified on the basis of the fact that

$$Q = f'(L V_F^{-1} L')^{-1} f \equiv (F - X b)' V_F^{-1} (F - X b)$$

where $b = (X' V_F^{-1} X)^{-1} X' V_F^{-1} F$ represent weighted least square estimates for the underlying parameters. In view of this IDENTITY and the large sample validity of the Wald Statistic Q, the weighted least squares estimates b are also regarded as having reasonable statistical properties because of the manner in which they determine Q. With these considerations in mind, it then can be noted that $V_b = (X' V_F^{-1} X)^{-1}$ represents a consistent estimate for the covariance matrix for b

When an appropriate model has been determined statistical tests of significance involving b may be performed by standard multiple regression procedures. Linear hypotheses are formulated as H_0: $Cb = 0$, where C is a known (d x u) coefficient matrix and tested using the statistic

$Q_C = b'C'[C(X' V_F^{-1} X)^{-1} C']^{-1} C\,b$ which is approximately distributed according to the χ^2-distribution with D.F.=d when the hypothesis H_0 is true.

Successive uses of the goodness of fit tests and the significance tests specified by the C matrices represent ways of partitioning the model components into specific sources of variance. In this context, the Q_C statistics reflect the amount by which the residual sum of squares goodness of fit Wald statistic Q would increase if the basic model were simplified (or reduced) by substitutions based on the additional constraints implied by H_0: $C\,b \hat{=} 0$. This partitioning of total variance into specific sources represents a statistically valid analysis of variance for estimator functions F arising from complex sample survey situations.

Finally, predicted values corresponding to any specific model can be calculated from $\hat{F} = X b = X (X'V_F^{-1} X)^{-1}X'V_F^{-1}F$ and corresponding estimates of variance can be obtained from the diagonal elements of $V_F = X(X'V_F^{-1}X)^{-1}X'$. Such predicted values not only have the advantage of characterizing essentially all the important features of the variation in the original data, but also represent better estimates than the original function statistics F since they are based on the data from the entire sample (ie., all subdomains combined) as opposed to its component parts. Finally, they are descriptively advantageous in the sense that they make trends more apparent and permit a clearer interpretation of the effects of the respective independent variables comprising X on the vector F.

3. APPLICATIONS AND EXAMPLES

3.1. Strategies for applying the model. In applying the methodology of Section 2 it is necessary to have a data analytic strategy. This strategy depends on taking note of the apparent bimodal nature of the statistical sciences. There are two general types of statistics -- descriptive statistics and inferential statistics. The role of the descriptive statistic is primarily summarization and does not strictly justify any comparative or other type of conclusion being extracted from the data. On the other hand, the inferential statistic is often a dimensionless index which may have only limited descriptive value. Its primary role is an orientation toward decision-making in the sense of being interpreted as either consistent with or in contradiction to a particular hypothesis which is of interest with respect to formulating conclusions from the data. Hence, certain inferential statistics can be used to determine whether any observed differences between two groups of individuals, such as a group of healthy patients and a group of diseased patients, are real or systematic as opposed to being due to chance variation; others can be used similarly to interpret the association among certain variables which are, for example, indicative of clinical status.

The preceding remarks have been directed at some of the objectives of statistical analysis. As formulated here, they appear reasonably clear and concise. However, the various ways in which statisticians operate in accomplishing them often appear heuristic and mystical. This impression results from the fact that "statistics" is in some sense an estranged marriage between "routine data processing" and "abstract mathematical probability theory." The paradox here is that "data processing" exists in the real world and can always be used to produce descriptive measures like arithmetic averages, percentiles, and least squares coefficients from any set of data, no matter how collected in terms of the underlying research design. Such computations constitute what will be called a "numerical analysis" of the data. Of course, the conclusions resulting from this type of approach are entirely limited to the data under consideration and cannot be rigorously generalized to any larger underlying population from which it is a sample. Moreover, a strict "numerical analysis" does not permit us to formally document the precision or reliability of quoted descriptive summary measures.

On the other hand, probability theory is a set of abstract axioms, definitions, and theorems, all of which are very much outside the real world, but which, under suitable assumptions, can provide reasonable mathematical models for characterizing numerically measured quantities. Within this framework, "statistics" is the liaison between a set of data and a suitable mathematical probability model. Hence, the most crucial aspect of any statistical analysis is the validity of the formal assumptions which underlie the corresponding probability model. Indeed, this principle can be underscored in some cases to the extent of identifying those assumptions or conditions which lead to contradictory conclusions and then choosing that conclusion together with its supporting analysis, for which the corresponding set of assumptions seems to be most empirically and/or physically realistic. It is in this context that the paradoxes associated with the sayings dealing with "how to lie with statistics" can be resolved.

Although the previously described point of view appears somewhat different from that which is concerned with developing standards for justifying statistical statements of a descriptive or inferential nature, there are definite similarities with respect to the underlying philosophy. In particular, the Q statistics in (2.8) represent an analytical procedure for assessing whether there is any variation among a given set of statistics and if so whether it can be partitioned in meaningful manner. The former question is entirely of an inferential nature and can be interpreted as dealing with the multiple comparison problem in a Scheffe simultaneous test procedure sense. The latter is both descriptive and inferential and may also be possibly guided to some extent by substantive considerations pertaining to the subject matter area for which such analysis is being undertaken. Similar remarks can be applied to the predicted values which are generated from various fitted models. These points will all be discussed in more detail for the examples in Section 3.2. In summary, there are two basic goals which are associated with the analysis of complex sample survey statistics as well as any other types of data:

1. Sample statistics which are essentially similar (in the sense of not being statistically different in a significance testing context) should not be reported in tables as different, although raw or unanalyzed data should of course be displayed where appropriate. Alternatively, the same estimate should be reported for each element in such cases. Moreover, one method for obtaining such estimates is weighted least squares as described in (2.8).
2. Sample statistics which are significantly different (at some appropriate level;eg., α=.05) should be reported in terms of correspondingly different estimates. However, attempts should be made to structurally characterize such differences in terms of models which can be fitted by weighted

least squares. In this latter context, it should be recognized that both significance testing inferential considerations as well as percent explained variation descriptive considerations are important.

3.2. Examples of the model. The preceding remarks will guide us in analyzing the following examples. All four are taken from the National Center for Health Statistics, Office of Statistical Methods, (unpublished manuscript). This, rather than the original sources indicated on the examples, was used to facilitate the computation of standard errors. In some cases, the standard errors were provided in the original sources, however in others it could only be computed with difficulty or not at all. In all cases sample correlations between the statistics were assumed to be zero. However, some preliminary investigation suggests that the Q-statistics are conservative in the case of positive equal correlation and anti-conservative in the case of negative equal sample correlation. That is, for positive correlation some differences will go undetected while for negative correlation nonexistent differences will appear.

Our first example analyzes the estimates of the proportion of dentulous persons, ages 18-79, needing to see a dentist prior to next regular visit, in various marital states. Our preliminary model saturates the variation space and the Q-statistic of 18.66 for total variation indicates that significant variation exists among marital states with respect to a χ^2(D.F.=4) distribution. This permits an examination of classes to identify the ones contributing the greatest variation. b_5 which corresponds to the difference between the separated and never married groups generates Q=16.55 which is significant even

Example 1

Comparisons Among Several Subdomains Within a Domain

Subdomain of White Adults	Estimated Proportion Needing to See a Dentist at an Early Date	Estimated Standard Error	Aggressive Model: Smoothed Predicted Values	Aggressive Model: Estimated Standard Errors	Conservative Model: Smoothed Predicted Values	Conservative Model: Estimated Standard Errors
Married	.378	.022	.389	.019	.366	.016
Widowed	.420	.049	.389	.019	.366	.016
Divorced	.426	.058	.389	.019	.366	.016
Separated	.627	.071	.627	.071	.627	.071
Never Married	.318	.027	.318	.027	.366	.016

Preliminary Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} .318 \\ .060 \\ .102 \\ .108 \\ .309 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	2.97
$b_3 \hat{=} 0$	1	3.32
$b_4 \hat{=} 0$	1	2.85
$b_5 \hat{=} 0$	1	16.55
$b_2 \hat{=} 0, b_3 \hat{=} 0, b_4 \hat{=} 0$	3	5.74
$b_2 - b_3 \hat{=} 0,\ b_2 - b_4 \hat{=} 0$	2	1.06
Total Variation	4	18.66

Final Models

Aggressive

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} .318 \\ .071 \\ .238 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	4.69
$b_3 \hat{=} 0$	1	10.45
$b_2 + b_3 \hat{=} 0$	1	16.55
Model: $b_2 \hat{=} 0, b_3 \hat{=} 0$	2	17.60
Residual GOF	2	1.
Total Variation	4	18.66

Conservative

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} .366 \\ .627 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_1 - b_2 \hat{=} 0$	1	12.91
Model: $b_2 \hat{=} 0$	1	12.92
Residual GOF	3	5.74
Total Variation	4	18.66

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports (Unpublished Preliminary Draft), June 1973.
Original Source: National Center for Health Statistics, Vital and Health Statistics, Series 11, Number 36, "Table 5. Percent of dentulous adults who should see a dentist at early date by marital status, race, and sex: United States, 1960-62," p. 14.

in the total variation space, that is with respect to a χ^2 (D.F.=4) distribution. Further, there is no difference among the remaining ever married groups Q=1.06, with respect to a χ^2 (D.F.=2) distribution. The difference between the never married and remaining ever married groups is somewhat more subtle. There are two approaches. The conservative model groups never married, divorced, widowed, and married into one group and distinguishes only the separated group. The residual Q=5.74 is nonsignificant with respect to a χ^2 (D.F.=3) distribution but may conceal an important component of variation as revealed by the aggressive model. The aggressive model has three groups, never married, separated, and other ever married. This model fits very well with a residual Q=1.06 which is nonsignificant even with respect to a χ^2 (D.F.=2) distribution. The smoothed values fulfill our goal of displaying differences only where they 'exist'; the question of which model is to be preferred is one that should be settled on substantive grounds.

Estimates of mean baby birthweight for various family income and mother's employment status categories provide our second example. The total variation, Q=22.83, in the preliminary model is significant with respect to a χ^2(D.F.=5) distribution, so as in example one, an effort to characterize the groups must be made. Parameters b_2 and b_3 compare the first to third and the second to third income groups and b_2 is individually significant. In fact, the Q that examines $b_2-2b_3\hat{=}0$ is only .05 so such a characterization of income is plausible. The variation associated with employment Q=14.24 is significant with respect to a χ^2(D.F.=3) distribution so these effects may also be further examined. Moreover, there is no interaction (employment status by income level), Q=1.00, with respect to a χ^2 (D.F.=2) distribution. These considerations lead to our final model and smoothed predicted birthweights which show only significant differences. This model has a small residual, Q=1.18, which is non-significant with respect to a χ^2(D.F.=3) distribution. It is also parsimonious in the sense that it contains only three parameters which permits a reduction in the standard errors of the

Example 2

Comparisons Among the Corresponding Subdomains of Two or More Domains

Domain: Family Income Level	Subdomain: Wife's Employment Status	Estimated Mean Baby Birthweight (grams)	Estimated Standard Error	Smoothed or Predicted Birthweight	Estimated Standard Error
$3000-4999	Employed	3230	23.46	3232	16.65
	Unemployed	3290	17.96	3293	14.35
$5000-6999	Employed	3280	22.11	3263	12.82
	Unemployed	3320	18.08	3323	10.47
$7000 and over	Employed	3280	21.15	3294	15.81
	Unemployed	3360	18.36	3354	14.55

Preliminary Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 3360 \\ -70 \\ -40 \\ -60 \\ -40 \\ -80 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	7.43
$b_3 \hat{=} 0$	1	2.41
$b_4 \hat{=} 0$	1	4.12
$b_5 \hat{=} 0$	1	1.96
$b_6 \hat{=} 0$	1	8.16
$b_4 \hat{=} 0, b_5 \hat{=} 0, b_6 \hat{=} 0$	3	14.24
$b_4-b_5 \hat{=} 0,\ b_4-b_6 \hat{=} 0$	2	1.00
$b_2-2b_3 \hat{=} 0$	1	.05
Total Variation	5	22.83

Final Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 3293 \\ 31 \\ -30 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	9.47
$b_3 \hat{=} 0$	1	13.24
Model: $b_2 \hat{=} 0, b_3 \hat{=} 0$	2	21.65
Residual GOF	3	1.18
Total Variation	5	22.83

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports(Unpublished Preliminary Draft), June 1973.

Original Source: National Center for Health Statistics, Vital and Health Statistics, Series 22, Number 8, "Table 7. Average Birth Weight, number of birth, and percent distribution by birth-weight intervals according to family income in 1962 and whether mother was employed during pregnancy; United States, 1963 legitimate live births," p. 19 and Appendix I p. 30.

smoothed cell means. In this model the employment status effect, Q=13.24, is significant at α=.05 in a Scheffe multiple comparison sense with respect to a χ^2 distribution with 5 degrees of freedom in the total variation space, or 3 degrees of freedom in the total employment status subspace, or 2 degrees of freedom in the total reduced model subspace. Similarly the income level effect Q=9.47 is significant at α=.10 with respect to a χ^2 distribution with 5 degrees of freedom, or 4 degrees of freedom which pertains to the total income subspace, or 2 degrees of freedom.

The third example of our sample survey model and the techniques guiding reduction is the estimated mean scores on the Block design subtest for boys and girls ages six to eleven. As before the total variation, Q=1719.36, is significant however we will focus on the sex differential since the age effect is a marked trend. The variation associated with sex, Q=30.49 is significant with respect to a χ^2(D.F.=6) distribution hence further analysis is warranted. The variation associated with the age by sex interaction, Q=10.05, is also significant at the α=.10 level with respect to a χ^2(D.F.=5) distribution. This interaction is not present when only the last 5 age groups are considered jointly, Q=1.88. These considerations produce our final model. Here we have an increasing score with age trend, Q=1377.49, an age x sex interaction term Q=28.53 which combines boys and girls only in the first age group. These effects are significant with respect to a χ^2(D.F.=5) and a χ^2(D.F.=1) distri-

Example 3

Comparisons Among the Corresponding Subdomains of Two or More Domains

Domain Age	Subdomain Sex	Estimated Mean Score on Block Design Subtest	Estimated Standard Error	Smoothed Predicted Value	Estimated Standard Error
6 years	Boys	5.8	0.27	5.7	0.18
	Girls	5.7	0.24	5.7	0.18
7 years	Boys	8.5	0.29	8.6	0.24
	Girls	7.3	0.25	7.2	0.22
8 years	Boys	12.0	0.39	11.8	0.30
	Girls	10.3	0.36	10.5	0.29
9 years	Boys	14.0	0.46	14.0	0.34
	Girls	12.6	0.42	12.6	0.33
10 years	Boys	18.2	0.63	18.6	0.44
	Girls	17.5	0.55	17.2	0.43
11 years	Boys	22.3	0.62	22.0	0.50
	Girls	20.1	0.82	20.6	0.52

Preliminary Model

$$\underset{\sim}{X} = \begin{bmatrix} 1&1&0&0&0&0&1&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0&0&0 \\ 1&0&1&0&0&0&0&1&0&0&0&0 \\ 1&0&1&0&0&0&0&0&0&0&0&0 \\ 1&0&0&1&0&0&0&0&1&0&0&0 \\ 1&0&0&1&0&0&0&0&0&0&0&0 \\ 1&0&0&0&1&0&0&0&0&1&0&0 \\ 1&0&0&0&1&0&0&0&0&0&0&0 \\ 1&0&0&0&0&1&0&0&0&0&1&0 \\ 1&0&0&0&0&1&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0&0&1 \\ 1&0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 20.1 \\ -14.4 \\ -12.8 \\ -9.8 \\ -7.5 \\ -2.6 \\ 0.1 \\ 1.2 \\ 1.7 \\ 1.4 \\ 0.7 \\ 2.2 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_7 \hat{=} 0$	1	0.08
$b_8 \hat{=} 0$	1	9.82
$b_9 \hat{=} 0$	1	10.26
$b_{10} \hat{=} 0$	1	5.05
$b_{11} \hat{=} 0$	1	0.70
$b_{12} \hat{=} 0$	1	4.58
$b_7 \hat{=} 0, b_8 \hat{=} 0, \ldots, b_{12} \hat{=} 0$	6	30.49
$b_7 - b_8 \hat{=} 0, \ldots, b_7 - b_{12} \hat{=} 0$	5	10.05
$b_8 - b_9 \hat{=} 0, b_8 - b_{10} \hat{=} 0,$ $b_8 - b_{11} \hat{=} 0, b_8 - b_{12} \hat{=} 0$	4	1.88
Total Variation	11	1719.36

Final Model

$$\underset{\sim}{X} = \begin{bmatrix} 1&1&0&0&0&0 \\ 1&1&0&0&0&0 \\ 1&0&1&0&0&1 \\ 1&0&1&0&0&0 \\ 1&0&0&1&0&1 \\ 1&0&0&1&0&0 \\ 1&0&0&0&1&1 \\ 1&0&0&0&1&0 \\ 1&0&0&0&0&1 \\ 1&0&0&0&0&0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 20.6 \\ -14.9 \\ -13.4 \\ -10.2 \\ -8.0 \\ -3.4 \\ 1.4 \end{bmatrix}$$

Statistical Tests	D.F.	Q
Age x Sex: $b_7 \hat{=} 0$	1	28.53
Age: $b_2 \hat{=} 0$, $b_3 \hat{=} 0$, $b_4 \hat{=} 0$, $b_5 \hat{=} 0$, $b_6 \hat{=} 0$	5	1377.49
Nonlinear Age: $b_2 - b_3 \hat{=} b_3 - b_4$, $b_3 - b_4 \hat{=} b_4 - b_5$, $b_4 - b_5 \hat{=} b_5 - b_6$, $b_5 - 2b_6 \hat{=} 0$	4	45.61
Model	6	1717.41
Residual	5	1.96
Total Variation	11	1719.36

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports (Unpublished Preliminary Draft), June 1973.
Original Source: National Center for Health Statistics, Vital and Health Statistics, Series 11, No. 107, "Table 3. Mean and standard deviation of raw scores on the Block Design subtest of the WISC by age and sex ... United States, 1963-65" p. 21, and "Table III. Sampling errors for average raw scores on the WISC Vocabulary and Block Design subtests by age, sex, ... United States, 1963-65," p. 38.

Example 4

Evaluation of Trend Effects

Midpoint of Income Class	Estimated Percent Needing Early Dental Visit	Estimated Standard Error	Aggressive Model: Smoothed Predicted Value	Aggressive Model: Estimated Standard Error	Regression Model: Smoothed Predicted Value	Regression Model: Estimated Standard Error
$1000	51.2	3.3	50.2	1.9	50.8	2.0
$3000	50.5	3.1	50.2	1.9	46.6	1.6
$5500	40.3	2.2	41.2	1.2	41.4	1.3
$8500	32.4	2.4	32.3	1.4	35.2	1.2
$15000	23.6	2.6	23.3	2.1	21.7	2.3

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 51.2 \\ -0.7 \\ -10.2 \\ -7.9 \\ -8.8 \end{bmatrix}$$

Preliminary Model

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	.02
$b_3 \hat{=} 0$	1	7.20
$b_4 \hat{=} 0$	1	5.89
$b_5 \hat{=} 0$	1	6.19
$b_2-b_3 \hat{=} 0, b_2-b_4 \hat{=} 0, b_2-b_5 \hat{=} 0,$	3	2.52
$b_3-b_4 \hat{=} 0, b_3-b_5 \hat{=} 0$	2	.164
Total Variation	4	69.45

Final Model

Aggressive Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 50.2 \\ -9.0 \end{bmatrix}$$

Statistical Tests	D.F.	Q
Model: $b_2 \hat{=} 0$	1	69.16
Residual	3	0.30
Total Variation	4	69.45

Regression Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5.5 \\ 1 & 8.5 \\ 1 & 15. \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 52.8 \\ -2.1 \end{bmatrix}$$

Statistical Tests	D.F.	Q
Model: $b_2 \hat{=} 0$	1	65.71
Residual	3	3.75
Total Variation	4	69.45

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports, (Unpublished Preliminary Draft), June 1973.
Original Source: National Center for Health Statistics, Vital and Health Statistics Series 11, No. 36, "Table 3. Percent of dentulous adults who should see dentist at early date, by family income, race, and sex: United States, 1960-62," p.13.

bution respectively. The age trend is non linear, Q=45.61, with respect to a χ^2(D.F.=4) distribution. That the final model does not obscure any important variation is seen from the residual Q=1.96, which is non-significant with respect to a χ^2(D.F.=5) distribution. The reported smoothed or predicted scores show all significant differences. Note that boys and girls are equal only at year 6 and boys score higher at later ages. Also there is a substantial reduction in standard errors for the smoothed values which has resulted from the final model's characterization or partioning of the total variation space.

Our last example uses estimates of the percentage of dentulous persons in various income classes needing an early visit to a dentist. The preliminary model displays sufficient total variation(Q=69.45) with respect to a χ^2(D.F.=4) distribution to indicate significant differences among the five income classes. However, the variation between the first two classes(Q=.02) indicates they are the same and the first difference occurs with the third class(Q=7.20). Moreover, contrasting the classes(Q=.164 and Q=2.52) reveals approximately equal declines in the percentages. These considerations lead to two possible models. The regression model uses the midpoint of the income classes and accounts for the significant variation(Q=65.71) with respect to a χ^2(D.F.=4) distribution, in a Scheffé multiple comparison sense. But it could be argued that on substantive grounds the first two income classes are the same and further the income groupings are essentially arbitrary. Given this the aggressive model would be proposed. Again, all significant variation(Q=69.16) is accounted for but groups that are probably alike are not separated.

ACKNOWLEDGMENTS

This research was in part supported by the National Institute of Health (Grants GM-70004 and HD-00371). The authors thank Dr. Monroe G. Sirken and Dr. Daniel G. Horvitz for discussions which were helpful in formulating this methodology.

REFERENCES

1. Cornfield, J., On Samples from Finite Populations. JASA-39.
2. Forthofer, R.N. and Koch, G.G., Analysis for Compounded Funtions of Categorical Data. Biocs-29.
3. Grizzle, J.E., Starmer, C.F., and Koch, G.G., Analysis Categorical Data Linear Models. Biocs-25.
4. Hansen, M.H., Hurwitz, W.N., and Bershad, M.A., Measurement Errors in Censuses and Surveys. I.S.I. Bulletin-38.
5. Horvitz, D.G. and Thompson, D.J., A Generalization of Sampling Without Replacement from a Finite Universe. JASA-47.
6. Kish, L. and Frankel, M.R., Balanced Repeated Replication for Standard Errors. JASA-65.
7. Koch, Gary G., An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications. JASA-68.
8. Koch, Gary G. and Lemeshow, S., An Application of Multivariate Analysis to Complex Sample Survey Data. JASA-67.
9. McCarthy, P.J., Pseudo-Replication: Half Samples. I.S.I. Review-37.
10. Wald, A., Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. Trans. of A.M.S.-54.
11. Wilk, M.B. and Kempthorne, O., Fixed, Mixed, Random Models in Analysis of Variance. JASA-50.

THE NEGATIVE INCOME TAX EXPERIMENT: COMMENCEMENT EXERCISES

Harold W. Watts, University of Wisconsin

Introduction

The announcement of the completion of the New Jersey Graduated Work Incentive Experiment made by the Assistant Secretary for Planning and Evaluation, in the Department of Health, Education and Welfare on December 20, 1973 represents an important stage in the maturation of a process begun more than six years ago. It is somewhat like the completion of a degree program that is marked by a commencement or graduation ceremony. An important goal has been attained, but its worth in terms of the ultimate objectives supposed to flow from it has yet to be established. The evidence produced, ranging from the large and complex data files to the refined summary indicators, is now launched on its own course, outside the protective custody of those responsible for its development. In an important sense a new experiment is just beginning--its conclusions will tell us how the evidence from this and other social experiments will have an impact on the sponsor (or "client"), the scholarly community, the general public, and ultimately on social policies.

In this paper the "developmental process" will be very briefly reviewed; then various, more speculative, remarks will be offered about the future path of the career of this and similar social experiments.

The Development of the Evidence

The "results" of the experiment, necessarily boiled down to a manageable set of numbers, inevitably seem out of proportion to the time and talent and, above all, money that has gone into their production. For this reason it is useful to review both the steps involved in producing the evidence, and a more complete list of the range of available products.

The design of the experiment--encompassing the original specification, sampling procedures, allocations to treatments, questionnaire development, and drafting of a "pseudo-law" in the form of rules of operation for the simulated negative income tax--was the result of combined efforts of the sponsoring agency (OEO), The Institute for Research on Poverty at Madison, and Mathematica, Inc. at Princeton. The effort involved economists, sociologists, lawyers, and statisticians in a genuinely joint enterprise. It would be wrong to claim that the process was smooth and painless. But the controversies were resolved, once with the help of outside "arbitration", and the project continued, more strenthened than scarred by the conflict. This experience suggests to me that real inter-disciplinary activity in an important project can be inspired by a clear and present danger of failure.

The collection of data involved, most importantly, an hour-long survey every quarter for the three years of operation at each site. In addition, various kinds of data were accumulated as a by-product of the operation of the experiment. The Urban Opinion Surveys division of Mathematica was responsible for this activity, working closely with staff from Madison when necessary. A great deal was learned about panel surveys in low-income areas from this experience.

Preparation of a data base for analysis is a third critical phase of the development of information. This includes coding, initial data entry, design and implementation of file systems, and a very large amount of "data cleaning". The objective, to get basic information from the questionnaires as close as possible to the fingertips of the researchers in an immediately manipulable form, sounds deceptively simple. Anyone who plans to carry out a social experiment should begin early and expect to work late on this part of the task. The primary function of coding and data processing were carried out by Mathematica, and the data cleaning was a joint effort.

The analysis--and this should be regarded as the "first exploitation" part of a hopefully continuing process of analysis--has been the main preoccupation during the past year. The largest share of this work has been carried on at Madison. Comparatively little co-ordination, aside from securing adequate coverage of subject areas, was sought in this pahse. Individual reseachers or small coalitions carried out separate analyses of specific topics. There were several suggested standardizations--sample specifications, definitions of key variables, etc.--which were widely adopted for reasons of convenience and also because of a desire for comparability with other analyses. More than 25 professional analysts were involved in this part of the job* along with at least as many support personnel.

The physical product of these efforts amounts to approximately 1500 pages of final report which are accompanied by an additional 450 pages of administrative procedures and findings which document the operations carried out by Mathematica, and draw conclusions from them.

In summary, then, there is a lengthy "final" report document which contains the complete set of analytic and descriptive studies produced by the project. There is also a summary document prepared for their own release by H.E.W., as well as less "official" summaries prepared for delivery at various meetings and seminars. Another output is the data file itself which is now available and accessible to the general research community for further analysis. A substantial effort is being made to inform interested scholars about the opportunity to use these data and to facilitate such use. The file is necessarily complex, particularly for those not accustomed to panel data, and we are attempting to short-cut many of the delays and frustrations that are typical when a researcher tries to exploit a "new" source of data.

The Evaluation of the Evidence

How will various "consumers" react to the

introduction of this large body of new and somewhat unprecedented evidence? At least three consumers can be distinguished: the immediate sponsors of the project and their policy-making adversaries or collaborators, the scholarly and research sub-cultures, and the public at large. Over the coming weeks and months each group will be exposed to some degree to this new evidence. It is of interest to consider how the evidence will be presented, perceived, and assimilated.

The primary recipient of the evidence in in the policy-making sphere is, of course, the sponsor. H.E.W., as the inheritor of the cognizant part of O.E.O., assumed responsibility for monitoring this experiment and for receiving and announcing the basic findings. As sponsor, H.E.W. received the complete range of detailed and summarized findings prior to general release and took a major role in the selection and interpretation of the evidence given emphasis in official releases and digests. The Office of Assistant Secretary for Planning and Evaluation serves as a distribution point both for other parts of H.E.W. and for other departments and agencies whose concerns touch on labor supply, taxation, or incentives.

The summary material has, in my opinion, been carefully and competently prepared. Close consultation with the experimental staff has been maintained and a judicious balance between the needs of brevity and accuracy has been achieved. These materials are of very great importance, because nearly all of the most immediately involved officials simply must rely on 5 to 10 page summaries. More lengthy documents can be digested by staff, of course, but this process again results in condensed versions for final consumption.

How much interest and effort is expended on the assimilation of the evidence in the administrative branch naturally depends on whether a legislative initiative is being developed to which the evidence is relevant. At the present time I have no particular insight about the likelihood of new administrative moves in the area of welfare reform, but there is, in any case, an evident continuing interest in this topic within H.E.W.

The official summaries and interpretations are important for another reason. If the evidence is used by the administration (or deemed a significant potential basis for contention), the staffs and services which inform the legislative branch will also take the official summaries, rather than the voluminous, fully hedged, qualified, and somewhat confusing "full report" as the starting point for their assessment, critique, or total rejection of the evidence. A great deal of active interest exists in the fiscal policy subcommittee of the Joint Economic Committee, headed by Rep. Martha Griffiths. Their truly massive study of the system of income maintenance and public assistance is aimed at major reform, and as such has been an eager consumer of all kinds of relevant evidence.

The congressional committees which have direct responsibility for Social Security, taxation, public assistance and any newly proposed reform, have a strong latent interest in reviewing any evidence from the experimental work, whether the evidence is used to support or oppose a particular proposal. Both Ways and Means and the Senate Committee on Finance were exposed to very preliminary evidence from the experiment, both in written form and in direct testimony, while considering the Family Assistance Program (FAP) when it was proposed in 1969. If major new legislation is introduced in the next few years there is little question that these committees will be intensely interested in the experimental evidence.

By no means all the policy action in the public welfare and labor supply areas is at the federal level. State and local governments are directly and primarily responsible for certain policies addressed by the experimental evidence. The channels by which the evidence is communicated to these levels is much less clear, however. Some state and municipal governments have already requested copies of the reports, and further requests are likely to be generated by stories in newspapers or the broadcast media.

The general public will certainly depend on the media for most of its information about the new evidence. Occasional stories have appeared in the past, reporting descriptively on the existence and objectives of the experiment, or reporting preliminary findings and interpretations. It seems likely that many stories will appear based on the "final" report. It is, of course, impossible to predict how faithfully or with how much perspective the evidence will be portrayed and perceived. But it seems highly unlikely that any reporter will find it worthwhile to peruse the entire array of primary studies. They will probably also rely on some ready-made summary or on a very partial sampling of the basic studies. It is very easy to be critical of almost any journalistic account of scientific or scholarly studies, but it is not at all easy to offer constructive alternatives when the basic subject matter is quite complex.

In the long run, of course, change or reinforcement of beliefs and myths about labor supply and motivation of the poor will depend on a broader process than the public press and broadcast system. A vigorous public policy debate, featuring the experimental evidence would accelerate its perception. A slower process would probably involve a "trickle-down" via opinion leaders and the educational process. In any case the content of the public's perception will depend on the evaluations and assessments by both scholars and policy-makers of the validity and ultimate relevance of the evidence. The immediate impact of the report on the beliefs of the general public is likely to be small in any case, and any eventual impact will depend both on the report and on the evaluation--pro and con--which it inspires.

This brings us around to where we are today--at a meeting of scholars and researchers whose analysis, irreverence and wisdom is needed to establish the limits of credibility of the experimental evidence. Having been produced by scholars and technical research professionals, the main report is of a style and form to interest and provoke others of the same breed. And it is from this community that the most severe and comprehensive review must come.

Previous reports at professional meetings and journal articles have communicated the basic design, scientific objectives, and preliminary findings to

the concerned professions. Interest in the experiment has always been great, but so far there has been a lack of substantive evidence and conclusions to either attack or defend. This lack will have been corrected by the end of these meetings (December 30th at 1:30 p.m. to be exact) and both summary and detailed reports are now available for review. In addition, as mentioned above, the data files used for the analysis, as well as the basic data files from which they were constructed, are accessible for researchers who wish to replicate, modify, or extend the analysis performed to date.

It is hoped that by now (some would say "finally") there are reports on the findings which can satisfy most of the interested professional-scholarly community. Papers--ranging from non-technical summaries and descriptive papers which can provide an authoritative introduction for the non-specialist to the detailed technical reports of greatest interest to the specialist in either the subject matter (e.g. labor supply) or methodology--are available.

It is now time for the profession at large to give this work the kind of careful, challenging and even nit-picking review which is necessary if this evidence is to receive or deserve wide credibility. It is possible that the entire undertaking has produced evidence eventually judged as worthless. It is more likely that some initial conclusions will be overturned and some others qualified more appropriately, so that in the end a more consistent and warranted set of conclusions will emerge.

At least two groups will be especially interested in the critical review process. Those who have contributed to the empirical literature on labor supply in the past, usually on the basis of non-experimental cross-sectional data, will be concerned with establishing some kind of rationalization for the similarity or lack of it between our results and theirs. Secondly, those who are responsible for the development of evidence from other experiments, presently at an earlier stage of maturity, will be eager to discover and learn from our mistakes and triumphs (if any) with a view to immediate application to their own efforts.

At present there are two organized efforts aimed at critical review. The Russell Sage Foundation has sponsored a project directed by Dr. Peter Rossi which will be especially concerned with the sociological methodology and findings from the experiment. The Brookings Institution has a project directed by Dr. Alice Rivlin and supported by the Ford Foundation which is charged with review of social experimentation in general and which will devote a major conference this spring, with "commissioned" papers, to the review of the urban negative tax experiment.

Major summary papers will be published in the Journal of Human Resources early in 1974, which is also expecting to receive and publish manuscripts concerned with evaluation and criticism of this social experiment. No doubt other journals will publish both original and critical studies based on the experiment as well.

The studies comprising the final report will eventually be published after further editorial and substantive revision. The latter is aimed primarily at closing some gaps of comparability across separate studies, and to press a bit further the resolution of some outstanding puzzles in the current report. This publication, which will require three volumes, will not be available until 1975 and will probably coincide with publication of numerous critical reviews.

Conclusion

After a lengthy period of development, a new and somewhat novel body of evidence is being introduced to the outside world. A new phase of the social-experimentation "experiment" is thus being entered by the earliest of the income maintenance experiments. Clearly the eventual policy pay-off of the endeavor depends on how that evidence is received, and how it holds up under stress. This part of the experiment is quite unstructured and uncontrolled, however, and is for that reason exciting and hazardous. But, with whatever degree of justice, a verdict will eventually be rendered--based on the perceived value of this and succeeding social experiments. Scholars and researchers have an important critical role to play in digesting and evaluating this new evidence and in judging the utility of social-experimental evidence more generally. It is now time for that effort to begin in earnest.

*Officially at a half-time level of effort, for the most part, but frequently working more than full time.

ISSUES IN THE ANALYSIS AND DESIGN OF THE EXPERIMENTAL PORTION OF THE HEALTH INSURANCE STUDY

Joseph P. Newhouse, The Rand Corporation

Introduction and Objectives

In this paper I discuss some selected issues in the design and analysis of the experimental portion of the Health Insurance Study.[1] The objectives, methods of procedure, and significance of the experimental portion of the Study are discussed in Newhouse (1974); those desiring a general overview of the project are referred to that paper. In this paper I first briefly review the objectives of the experiment for those who do not wish to read the longer paper. I then discuss two problems which will have to be faced when analyzing the data from the experiment, and the implications those problems have for the design. In line with the chairman's charge to discuss "methodological questions, issues, and constraints," I have chosen problems whose solution will require breaking new ground; in neither case do I feel we have reached a definitive answer. I conclude by discussing three statistical design problems we have had to solve.

Stated at the most general level, the objective of the experiment is to advance the state of knowledge concerning the consequences of alternative ways of financing medical care services. We seek to measure own- and cross-price elasticities (insurance elasticity) of demand and their interactions with income. Measurement of price elasticity is a necessary condition for predicting utilization and cost under any particular insurance plan, and if the supply of services is perfectly elastic in the long run, it is sufficient to predict how insurance will affect the share of the nation's resources devoted to medical care. Measurement of the interaction of price elasticity with income will determine the distributional effects of any particular financing plan. In those plans which require out-of-pocket payments on the part of the consumer, we have designed the plan so as to limit his maximum out-of-pocket loss to a certain percentage of his income, since this is a potential policy option. The maximum out-of-pocket loss is called the Maximum Dollar Expenditure (MDE). Thus, price elasticity is to be measured within the context of this type of plan.

We also seek to measure, as best we can, the effects of alternative financing arrangements on health status. Whereas our first goal, the measurement of price elasticities, may be thought of as relating to the costs of a financing method, this second goal can be thought of as measuring certain of the benefits.

A third goal relates to understanding the consequences of increasing the demand for ambulatory services. Analysis shows that a national health insurance plan could cause a substantial disequilibrium in the market for outpatient physician services. This in turn could lead to the activation of several kinds of mechanisms to equilibrate the market, including price increases, queuing, delays to appointments, change in case-mix seen by physicians, changes in revisit rates, and so forth. The extent to which each of these mechanisms operates will play an important role in determining who gets what kind of service for what kind of medical problem. We seek to provide some information on how the burden of adjustment is distributed among these mechanisms.

A fourth goal is to measure the effect of prepaying the physician for his services rather than paying him on the basis of fee-for-service. A fifth goal is to find out how much additional private insurance families would buy if there were a public plan which required out-of-pocket payments for services (as Medicare does, for example). Finally, we wish to learn as much as possible about the administrative problems and rules of operation which arise in health insurance plans, particularly those which have income-related clauses.

In order to estimate the effect of price on utilization and health status, we have structured an experiment which will give various health insurance plans to approximately 7,500 individuals (in 2,000 families) in four sites.[2] The insurance plans are structured so that the families pay a percentage of their bill which varies from zero to 100 percent. As mentioned above, if the family must pay something out-of-pocket, its expenditures are limited to a certain fraction of its income; the fraction varies as an experimental treatment; it is either five or fifteen percent. In some other plans all outpatient care is free, but the family must pay a specified fraction of inpatient expenditures. Also, some individuals are to be enrolled in a Health Maintenance Organization, in which the physicians are prepaid.

Observations on the utilization of the participants should establish the price elasticity of demand, as well as the effect of prepaying the physicians. The effect of insurance on health status is extremely difficult to assess because of the difficulty of measuring health status. In order to measure self-assessed health status, all of the participants will take quarterly interviews; all of them will also take screening type physical examinations at the end of the experiment to measure "objective" health status; some participants will take initial physical examinations.[3]

Measurement of the consequences of a disequilibrium in the market for ambulatory services is accomplished by selecting sites in which the physicians' workload varies. While the range of variation in workload across communities may not include the workload which would be observed if free ambulatory coverage were instituted, it is the only method within the context of the experiment to obtain information on this important question.

The degree to which families will supplement not very generous insurance will be measured by permitting supplementation in the final year of the experiment. By that time we will have an estimate of the actuarial value of the policy; we intend to offer supplementary insurance at varying rates in order to test the effect of alternative tax treatment of health insurance premiums. Such premiums are not now taxable income if paid by the employer, but it has been proposed that this treatment be changed.

Some Issues in Analyzing Data from the Experiment

A principal issue which the analyst of the experimental data will face is the treatment of price, given that price falls with total expendi-

ture because of the MDE. As a result, traditional methods of analysis are inappropriate. Prior work in the field of medical demand analysis and demand analysis more generally has tended to analyze consumption (either measured in dollars or in physical units) per unit of time as a function of price. In these analyses price per unit is assumed to be constant. The theory underlying these analyses is standard economic theory, which assumes that the consumer optimizes, such that he values the marginal unit at the marginal utility of income foregone to purchase it. With the MDE, however, there are two local optima, as shown in Fig. 1. Fig. 1 shows a two-commodity world of medical care and all other goods; I_1 and I_0 are indifference curves.[4] The kinked line is the budget line; after the consumer has consumed L units of medical care, he does not have to sacrifice more of other goods to obtain care. (The budget line is net of any taxes of premiums the consumer has paid to finance the insurance policy.) There are two local maxima, at A and B; in this diagram B is clearly the global maximum.

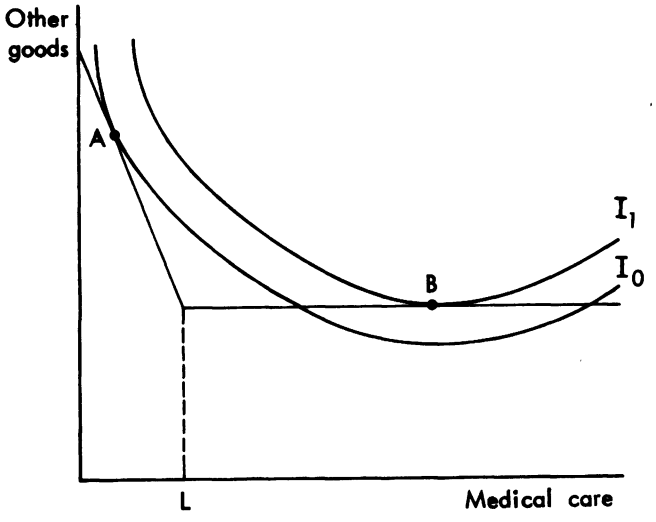


Figure 1

The problem caused by the kinked budget line, while somewhat novel from the point of view of empirical demand analysis, is reasonably tractable. Consumers can be assumed to have a utility function of a specified kind; then by observing expenditure choices, one can infer the parameters of the function.[5] Knowing the function, one can predict the consequences of any price structure.

What makes experimental data difficult to analyze is that the consumer will typically face several choices during the expenditure accounting period, none of which taken singly could cause him to exceed the deductible, but all of which together may. Thus, when making his initial choices, the consumer is operating under uncertainty about what the marginal price at the end of the accounting period will be.

Our plan for modeling this situation is to associate all expenditures with an illness episode.[6] We try to explain an individual's expenditures on each illness episode using as explanatory variables the insurance plan the consumer has, the consumer's expectations regarding expenditures on this episode and future expenditures, the amount of expenditure the consumer must make before his coinsurance rate changes, and a set of demographic variables such as income. We separate expected expenditures on this episode and expected expenditures on future episodes, because the consumer has information about this episode when he begins therapy which he does not have about future episodes. Expenditures on this episode are a function of diagnosis (using extraneous data),[7] and expenditures on future episodes are a function of the individual's age, sex, general health status (including any chronic conditions), and so forth.[8] The theory underlying an episodic model is discussed at length in Phelps (1973), although Phelps does not treat the case of a price which falls with expenditure.

The resulting equation (together with an error term) generates a distribution of expenditure per episode per individual; there is also a distribution of episodes which the experimental data will generate. If these distributions are analytically tractable, they can be convoluted, and mean individual expenditure per year (or other accounting period) predicted. If they are not analytically tractable, a simultion can be performed to predict mean expenditure per year. If there are family related clauses in the insurance plan (for example, family deductibles), individual expenditures must be aggregated to the level of the family.

The problem of price per unit falling with quantity and the proposed solution of analysis by episode have several implications for the design. The most obvious is that data must be gathered which permit us to define illness episodes and link medical services to them. This implies a degree of cooperation on the part of the physicians in filling out claims forms; it also implies assuring ourselves that the application of the definition of an episode is reliable. A second implication is to minimize the number of discontinuities or kinks in the price line in order to simplify as much as possible modeling the uncertainty which the consumer faces. This has been done by limiting all plans to at most one change of price; that is, there will be one coinsurance rate which will apply to family expenditures until expenditures have reached a certain percentage of family income, after which there is no coinsurance. (There is no plan with a deductible followed by a non-zero coinsurance rate, followed by some other limit.) A third implication is to stipulate that there be no carryover of unreimbursed expenditures from one accounting period to another. At one extreme there could have been a moving average accounting period; this would mean there would be no coinsurance so long as the rate of expenditures exceeded a certain amount. Besides creating a perverse incentive to consume services (unless based on the rate of unreimbursed expenditure), this arrangement would be extremely difficult to analyze. A modification of this, which was considered at some length, was to permit carryover of unreimbursed expenditures occurring at the end of the accounting period to the next accounting period. The intent of this is to prevent someone from having to satisfy an expenditure limit twice in the same illness episode, which could happen if the episode occurred near the end of a fixed accounting period. While conceptually straightforward, it was felt that the empirical complications such a clause would introduce were

not worth the value of having it.

The final implication of episode analysis for the design is that the design should provide a hedge, if episodic analysis proves infeasible. The hedge is that an analysis of covariance model (basically estimation of means across plans adjusted for demographic differences) should yield reasonably precise estimates of a few dissimilar plans. This can be done by minimizing the number of plans. There are now 16 plans and around 8,000 family years to be allocated to them. Our estimates are that this should yield estimates of the effects of insurance to within plus or minus 10 percent or so (Newhouse 1974). If, however, episodic analysis proves infeasible, sequential design will permit even greater concentration of families among a smaller number of plans. The current time schedule calls for fifty families to begin as a pilot sample on January 1, 1974. If all goes well, enrollment of the next 500 families will take place in late summer of 1974, the next 500 families nine months later, the next 500 families six months later, and the last 500 families three months after that. As a result, concentration of families in a few plans will be feasible should it prove necessary.

A second set of analytical issues arises in connection with supplementary behavior. These issues concern the design and analysis of the supplemental portion of the experiment. At what terms should supplementary insurance be available? Should both positive and negative supplementation be permitted? Should an individual be allowed to vary his coinsurance rate, his expenditure limit, or both? What is the dependent variable for analysis and what are the explanatory variables? We next discuss these questions in turn.

Since the overall objective is to understand the demand for supplementary insurance, the terms at which such insurance can be purchased will reflect the rates at which it would be available in the marketplace, if a national plan were enacted which resembled our plan. Therefore supplementation will be permitted at different premiums, reflecting both the tax treatment of insurance premiums and various loading charges. Insurance premiums are not now taxable income if paid by the employer, which means there is a subsidy equal to the marginal tax rate from the purchase of insurance. Loading charges vary from 100 percent for some individual insurance to 6 percent for individuals in the largest group. We will offer insurance at a small number of loading charges, ranging from 100 percent down to minus 30 percent or so, reflecting a tax subsidy and a very low loading. The actuarial value will be adjusted for the age and sex mix of the family, since the private market is likely to take account of such differences.

No decision has been made on whether both positive and negative supplementation will be allowed. If negative supplementation were not allowed, an individual could purchase more generous insurance, but not less generous.[9] To allow positive supplementation is sufficient to provide data on the degree to which individuals would purchase additional insurance if the government mandated a particular plan. To allow both positive and negative supplementation is to attempt a broader study of the demand for insurance. If negative supplementation is allowed, nothing is sacrificed in terms of measuring what would happen if only positive supplementation were allowed (since anyone who negatively supplements would just be assumed not to supplement if such were not allowed). However, negative supplementation raises an ethical and a practical issue, because allowing negative supplementation on the basis of actuarial value raises the possibility that an individual could be ex post worse off.[10] (For example, an individual with full coverage who chose to change to $500 deductible and received, say, $200 representing the actuarial value of the difference, could have a bill of $500.) If an individual were to be worse off ex post, it would usually be to his advantage to withdraw from the experiment and return to his old insurance. Current HEW guidelines on research with human subjects require that withdrawal be permitted. Withdrawal under these conditions would obviously defeat the experiment. In order to prevent withdrawal, one can make lump sum payments to those who have generous insurance plans in an amount large enough to compensate them for their worst case.[11] Whether this is sufficiently expensive so as not to be worthwhile depends upon the actuarial values involved; the higher the actuarial value, the less must be paid for worst case compensation. As noted above, no decision has yet been made on negative supplementation.

Given that positive supplementation at least will be allowed at various loading charges, the question arises as to what kind of supplementary behavior should be permitted. The tentative answer to this question is that individuals should be permitted to choose coinsurance rates of 0, 25, or 50 percent, and that they should be allowed to set their MDE at 5 percent of income. (The zero coinsurance rate is equivalent to a zero MDE.) The basic reason for allowing variation in both dimensions (coinsurance and MDE) is that the private insurance market would offer such variation, and it is important to understand how much each dimension is varied, since there may be different implications for demand, according to which dimension is varied.

Choice of explanatory variables will be based on theoretical work related to consumer choice of insurance; unfortunately, this literature is not yet very far advanced.[12] Nevertheless, it is clear from the work which has been done that the consumer's choice will be a function of the distribution of the consumer's expected expenditures, his permanent income, and the price of supplementation. Depending upon these variables, the consumer chooses an optimal MDE-coinsurance pair from the pairs that are open to him. Each of the two dimensions of choice will be analyzed separately. While the choices in each dimension are interdependent, theory is not powerful enough to specify restrictions which exclude a variable from one equation and not the other. Hence, structural equations are not identified and only a reduced form equation will be estimated. Since the choice of supplementary insurance which we have structured is discrete, methods developed by Nerlove and Press (1973) for estimation with polytomous dependent variables will be used.

An alternative to analyzing choice of MDE and coinsurance is to analyze the risk which the consumer leaves himself bearing. Provided suitable

measures of risk can be found, this offers the possibility of testing hypotheses related to risk-bearing behavior. Analysis of risk per se, however, is not sufficient for policy purposes, since there may not be any convenient way of moving from a measure of risk to a unique structure of insurance. The structure is important for policy purposes, since it will affect demand (except in special cases). Even if there were a way to move from risk to a structure, it appears more efficient to work directly with the structure, if that is what one is interested in. A similar argument can be made for not measuring supplementation by the amount of the supplementary premium that the consumer pays.

Issues in Statistical Design

Among the many possible issues in the statistical design of the experiment, four will be discussed here. These are the choice of the number of individuals to be assigned to any particular plan, selection of participating individuals from the community, allocation of those individuals to plans, and choice of sites in which to experiment.

The number of individuals to be assigned to any plan will be determined by use of the Conlisk Watts model developed for the New Jersey Negative Income Tax Experiment (Conlisk-Watts 1969). This model assumes that one is interested in estimating a vector of coefficients β in a model:

$$y = X\beta + \varepsilon, \text{ where } E(\varepsilon) = 0 \text{ and } V(\varepsilon) = \sigma^2 1. \quad (1)$$

The estimator of β is:

$$b = (X'X)^{-1} X'y, \; V(b) = \sigma^2 (X'X)^{-1}.$$

The admissible regressor rows of X are specified and consist (in our case) of prices (as determined by insurance plans) or a set of dummy variables for insurance plans (in an analysis of covariance model). A budget constraint is assumed and cost per regressor row (insurance plan) is given. The model then chooses the number of observations to be assigned to any design point such that

$$\phi = tr(WV(Pb)) \text{ is minimized,} \quad (2)$$

where W is any vector of weights and P is any arbitrary set of vectors, but most frequently equals either I or X.

Apart from specifying W and P, the major issue to be resolved in this step is the choice of X. Within this issue there are in turn two subissues. One is the choice of plans, or equivalently, the admissible rows of X. Since the model is free to allocate no observations to any design point, this choice really concerns which design points are constrained to have no observations assigned to them. The second is the issue of what functional form to choose.

Choice of design points can be thought of as first determining the number of design points which are not constrained to be zero (that is, determining the maximum number of insurance plans) and then determining what those design points are. While constraining fewer points to be zero will lead in general to a smaller value of ϕ, there are two costs to considering larger numbers of design points. The first is a computational cost. The second follows from the desire to hedge, discussed above, by maintaining the viability of the analysis of covariance model. More design points degrade the precision of this model. As mentioned above, there are initially sixteen design points; they are described in Newhouse (1974).

Several possible functional forms will be considered. It is envisioned that a decision-theoretic approach to choice of functional form will be used, so that functional form will be chosen which minimizes expected loss. (See Conlisk-unpublished)

A secondary issue related to the design of plans is the possibility of truncating the MDE in order to achieve greater efficiency in the estimation of price elasticities. Truncation may represent a gain because the method of reimbursing families generally requires that they receive a lump sum equal to their MDE. The question then arises: What is the value of an absolute dollar ceiling on the MDE? This has the effect of eliminating price variation in certain high ranges of expenditure, while increasing the budget available to allocate to lower ranges of expenditure. The optimal amount of truncation therefore depends on the relative interest in the estimation of elasticities at different levels of expenditures. We have structured this problem so that the Conlisk-Watts model may be applied to it. Rows of the X matrix will be created representing plan-total expenditure pairs. We then associate with each plan-total expenditure pair the conditional probability of that observation, given that the individual is assigned to that plan. The constraint is then placed on selection of regressor rows that all rows associated with a particular plan must be selected if any are. The MDE (together with any truncation factor) enters the model as changing the costs of an insurance plan as well as the probabilities of obtaining expenditures in various intervals. By running the model with alternative MDEs (and given weights for expenditure-elasticity interactions), one obtains a set of truncations which minimizes ϕ. For example, the optimum might be the lesser of 15 percent of income or $1,200 in the 100 percent coinsurance plans, but the lesser of 15 percent of income or $600 in the 25 percent coinsurance plan.

Individuals will be chosen for this experiment by means of the Finite Selection Model (FSM) developed by Morris for this experiment (Morris, forthcoming (a)). This model is similar to the Conlisk-Watts model in its objectives, but quite different in its underlying assumptions. The Conlisk-Watts model assumes that the regressor rows (the rows of the X Matrix) come from discrete space, but that there is an infinite population to sample from. This assumption is appropriate for selection of treatments, but much less so for selection of families. By contrast, the FSM assumes that the regressor rows may come from continuous space, but that there is a finite population to sample from. For example, individuals have associated with them a vector of demographic characteristics which are continuous or nearly so (age, family, size, income, education, for example). There are, however, only a finite number of possible individuals to choose to participate.

More formally, the FSM assumes that one is interested in estimating equation (1) from a subset of size n of all N available families, $n < N$. The objective function is given by (2). If there are n observations, (2) can be rewritten as:

$$\phi_n = \text{tr } D(X'_n X_n)^{-1} = \text{tr} DS_n \ , \qquad (3)$$

where the subscript n indicates that there are n rows in X, and D equals P'WP. Suppose an $(n+1)^{st}$ observation is to be added which reduces ϕ_n as much as possible for its cost. That is, we wish to maximize

$$(\phi_n - \phi_{n+1}(x))/c(x) \ , \qquad (4)$$

where ϕ_{n+1} is conditional on using an X matrix equal to $\begin{matrix} X_n \\ x' \end{matrix}$ and x' is a row vector of characteristics of the $(n+1)^{st}$ family. c(x) is the cost of a family with characteristics x. An algebraic identity gives

$$S_{n+1} = S_n - \frac{S_n x x' S_n}{1+x'S_n x}$$

Hence,

$$\phi_{n+1}(x) = \text{tr} DS_{n+1} = \text{tr} DS_n - \frac{\text{tr} DS_n x x' S_n}{1+x'S_n x}$$

$$= \phi_n - \frac{\text{tr } x'S_n DS_n x}{1+x'S_n x} = \phi_n - \frac{x'S_n DS_n x}{1+x'S_n x} \ ,$$

since $x'S_n DS_n x$ and $x'S_n x$ are scalars.

Substituting, (4) equals:

$$\frac{x'S_n DS_n x}{c(x)\ [1+x'S_n x]} \ .$$

Given a list of unchosen individuals, (5) may be computed for each x and the maximizing x chosen. The procedure can be repeated until n is such that

$$\sum_{i=1}^{n} c_i = C,$$

where C is a budget constraint.

The stepwise algorithm implied by the successive use of (5) has in our experience led to an optimal, or nearly optimal, set of families, but did it not do so, substitutions and corrections could be applied at the end of the selection by using a similar algorithm until a satisfactory list is obtained. While the literature contains no discussion of the use of this algorithm on the ϕ objective function in this context of costs and variances, there is experience and theory for using a related algorithm to determine "D-optimum" subsets (choosing subjects to maximize det $|X'_n X_n|$), and the experience there has been favorable (Harville (1973), Johnson (1973), Mitchell (1973), Wynn (1972)).

While the FSM will be used to choose the families which will participate in the experiment, we will "randomly" allocate the families to experimental treatments (plans) and the control group.[13] While the FSM can, in principle, select optimal family-plan combinations (given a model to be estimated), random allocation offers some protection against latent variables. (That is, one can be reasonably sure that any such variable will be balanced among the treatment groups.) By allocating randomly one pays a price in efficiency of estimation (if there are no latent variables). The price paid can be kept small if random allocation is made subject to a constraint of near orthogonality between the demographic and plan variables, ensuring near balance among the treatment groups.[13]

The Conlisk-Watts model and the FSM may also be applied to site selection. Morris has proposed a generalization of the Conlisk-Watts model which can be used to determine the optimal number of sites, given that there are fixed costs of operating in each site (Morris, forthcoming (b)). We assume a random effects model for city-specific coefficients $\beta_i \sim N_k(\beta, T)$, where k is the dimensionality of the β vector. T is therefore the between-city variance-covariance matrix. For simplicity assume that the same design points are to be used in each city and that the cost of a design point does not depend on city. The model then minimizes:

$$V = \text{between site variance} + \text{within site variance} = (1/K)(v + \phi), \qquad (6)$$

subject to

$$C = K(C_0 + \sum_{j=1}^{m} c_j n_j) \ , \qquad (7)$$

where there are K sites, $v = \text{tr}(DT/\sigma^2)$, ϕ is defined by (2) for the observations within each site, C_0 is the fixed cost of operating in any site (for opening an office, running a field staff, and so forth), c_j is the cost of an observation at the j^{th} design point (or insurance plan) and n_j represents the number of observations at the jth design point in each city.

The solution to this problem is quite simple. Define $C_K = (C_T/K) - C_0$ as the amount per city available to spend on design points after paying the fixed costs per site C_0. Then, following the standard Conlisk-Watts procedure, minimize (2) subject to a budget constraint of C_K. This defines $\phi^{(K)}$, where the superscript indicates that ϕ depends on K. The optimal K is that integer K* which minimizes $(1/K)(v + \phi^{(K)})$ and can be determined by enumeration in our case.

After determining the optimal number of cities in this fashion, we plan to use the FSM to select actual cities. The variable of interest across cities is the workload of physicians; the FSM will tend to select extreme values. A city-specific cost index will also be entered in the FSM, so that it will tend to select cheaper cities. The first site is Dayton, Ohio; the second site will be in the West and the third site in the South. No decision has been made on the location of the fourth site.

NOTES

1. The research reported herein was performed pursuant to a grant from the Office of Economic Opportunity, Washington, D.C. The opinions and conclusions expressed herein are solely those of the author and should not be construed as representing the opinions or policy of any agency of the United States Govern-

ment. The author would like to acknowledge the assistance of Carl Morris on the issues discussed in the last part of the paper.

2. There is also a control group who remain on their existing insurance. Their purpose is described in Newhouse (1974).

3. Some of the participants will not receive an initial physical in order to measure the effect, if any, of the physical on utilization.

4. The only unusual thing about them is that they turn up, indicating that the consumer has negative value for certain levels of medical care consumption. This may be because it takes increasing amounts of time (assumed to have increasing value) to consume more care or because sufficient exposure to medical care may actually decrease health status, through increasing the risk of infection or iatrogenic disease, relative to the possible benefits of care.

5. This suggestion has been made by Kenneth Arrow.

6. There are two types of episodes. An acute episode represents the consumer's response to a random loss of health stock; an acute episode in general will terminate within a relatively short period of time, either because the underlying pathology is self-limiting or because medical intervention has cured the problem. A second type of episode is chronic; a chronic problem in general requires medical intervention to maintain a stock of health and is not expected to terminate. The chronic episode therefore lasts for the entire accounting period. (A chronic condition in remission which "flares up" will be treated as an acute episode.) For analytical purposes it differs from an acute condition in that expenditures may be assumed to be better foreseen.

7. We will estimate expected expenditure by using mean expenditure for that diagnosis (if possible mean expenditure conditional on a particular plan). This implies we must measure the incidence of episodes for which no care was sought, which we will attempt to do in quarterly interviews.

8. If possible, generalized least squares will be used to allow for non-zero covariances among family members.

9. For example, an individual with a policy with a $1000 deductible who was being paid $1000 could change the deductible to $500 and be paid $800, but an individual with full coverage would not be allowed to choose a deductible of $500 plus $200, if that were the actuarial value of the difference between a full coverage policy and one with a $500 deductible.

10. To allow negative supplementation of not generous plans also raises the issue of accurate calculation of actuarial values; we will not be well placed to determine, say, the actuarial value of increasing a deductible from $1000 to $2000.

11. These payments may be held in an escrow account and made conditional on completion of the experiment.

12. Theoretical beginnings may be found in Arrow (1973a, 1973b), and Phelps (1973).

13. The idea contained in this paragraph was suggested by Bradley Efron.

REFERENCES

Arrow, Kenneth J., "Optimal Insurance and Generalized Deductibles" (Santa Monica, California: The Rand Corporation, R-1108-OEO), 1973a.

Arrow, Kenneth J., *Welfare Analysis of Changes in Health Coinsurance Rates* (Santa Monica, California: The Rand Corporation, R-1281-OEO), 1973b.

Conlisk, John, "Choice of Response Functional Form in Designing Subsidy Experiments" (unpublished mimeo).

Conlisk, John and Harold Watts, "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," *Proceedings of the Social Statistics Section*, American Statistical Association, 1969, pp. 150-156.

Harville, D. A., "Computing Optimum Designs for Covariance Models," *International Symposium on Statistical Design and Linear Models*, Colorado State University (Fort Collins, Colorado), March 19-23, 1973 (mimeo).

Johnson, J. D. and K. W. Last, *Constrained Experimental Design*, Lockheed Palo Alto Research Laboratory (Palo Alto, California), (mimeo).

Mitchell, T. J., *Applications of An Algorithm for the Construction of "D-Optimal" Experimental Designs in N Runs*, Statistics Department, Oak Ridge National Laboratory, March 1973 (mimeo).

Morris, Carl, "A Finite Selection Model for Experimental Design" (Santa Monica, California: The Rand Corporation), forthcoming(a).

Morris, Carl, "Determination of the Number of Sites in the Housing Allowance Demand Experiment" (Santa Monica, California: The Rand Corporation, R-1082-HUD), forthcoming(b).

Nerlove, Marc and S. James Press, "Univariate and Multivariate Log-Linear and Logistic Models" (Santa Monica, California: The Rand Corporation, R-1306-EDA), December 1973.

Newhouse, Joseph P., "A Design for A Health Insurance Experiment," *Inquiry*, March 1974.

Phelps, Charles E., "The Demand for Health Insurance: A Theoretical and Empirical Investigation" (Santa Monica, California: The Rand Corporation, R-1054-OEO), July 1973.

Wynn, H. P., "Results in the Theory and Construction of D-Optimum Experimental Designs," *Journal of the Royal Statistical Society, B*, Vol. 34, No. 2, 1972, pp. 133-147.

SOCIAL SCIENCE RESEARCH COUNCIL PROJECT ON SOCIAL EXPERIMENTATION

Henry W. Riecken, University of Pennsylvania

The purpose of this paper is to put forward an argument in favor of social experimentation as a method for planning and evaluating social interventions. This general position is the one that has been adopted in a monograph on Social Experimentation that has been prepared by a Committee of the Social Science Research Council and is now in press.

The argument in favor of social experimentation begins from a consideration of the inherent disadvantages of post-hoc "program evaluations". It is clear that program evaluation has recently become a popular activity for social scientists. Much recent social legislation includes a requirement for evaluation of the legislative program. The wave of domestic social reforms in the 1960's that led to compensatory education, community action programs, manpower training, and measures for diminishing racial segregation and sexual discrimination has been responsible for the creation of a mini-industry of evaluation. It is premature to judge how influential such evaluations have been in reshaping social policy, but experience to date suggests there are certain difficulties associated with the usual and ordinary procedure of conducting evaluations of national programs after the fact - that is, waiting until the program is put into full operation before giving appreciable attention to its evaluation. Many of the post-hoc program evaluations that have been carried out on national programs have produced negative evidence or evidence that there has been no change in the state of affairs the program was designed to alter.

This is a persistent difficulty with post-hoc evaluations for a variety of reasons. Many programs, for example, simply do not contain enough variations either in type of treatment; or in range of treatment intensity; or in characteristics of the units affected by the treatment to allow general conclusions to be drawn about alternative treatments. Where programs do contain such variations, treatment effects may be confounded by non-random assignment of the recipients of the treatment. For example, personal characteristics or motivations that lead people to volunteer for a program or to be first in line for a particular treatment may interact with the treatment effects. These are problems of experimental design.

In addition to such difficulties, there are management problems. The very attitude implied by post-hoc evaluations exacerbates certain managerial and institutional frictions. The evaluator turns up after the fact, so to speak, and presumes to make judgments about how well the individuals running the action program have done. The evaluator usually winds up telling them what they should have done if only they had been smarter in the beginning. Such advice often provokes resistance to it on the part of the program operators, who may correctly suspect that the evaluator's hindsight is keener than his foresight. Accordingly one of the management problems of post-hoc evaluation seems to be that the reports written by the consultants who have come in after the event, are received by those who are planning future programs with vast indifference or even hostility.

A final reason for questioning the effectiveness of post-hoc evaluation is that programs, once started, are very difficult to change. Both the clients of the program, and the managers of it get a stake in it. They have a stake in doing things in comfortable and familiar ways and are not very likly to welcome radical innovations, particularly those coming from outside. This is the unreceptivity phenomenon again, though for a different reason.

All four of these considerations suggests that one ought to adopt a somewhat different stance towards social intervention and its evaluation. Instead of establishing a program and then evaluating it, one ought to look at the matter as a cycle of program development, experimental test, and revision prior to installation on a larger scale. The cycle begins with an idea, a notion about how to intervene in a social process, which must then be developed into a program or treatment. To take a concrete case, the idea has been widely circulated in the medical and public health community that a protein-deficient diet of post-weaning children in many less developed countries is responsible for a certain amount of intellectual deficiency at school age and in adulthood. This opinion has resulted in proposals to feed protein supplements from six months of age (or whenever the weaning process is completed) up til school entrance age, when presumably, a substantial amount of physical growth and development has taken place. In order to test this proposition in an experiment it is necessary to develop a feeding program. That is, one must work out a dietary supplement which is acceptable, palatable, and protein-nutritious. One must develop some sort of system for administering the supplement, for making sure that the children who need it get it, that it is not sold in the local market by the families to whom it is given, and that it is not consumed by the adults, but indeed gets to the children who need it. All of these features of a program sound simple once they have been worked out, but they all need to be invented and made a functioning part of the treatment. Incidentally, in the course of developing an experimental treatment from an idea into an operating scheme, a good deal can be learned about potential problems and desirable administrative features of a full-scale program.

Following the development of treatments, the experiment itself must be designed. Since this is an audience of professionals, who know what experimental design is all about, there is no need to go into details about random assignment to treatments, protecting controls from contamination, etc. It should be

emphasized, however, that the SSRC monograph adopts a rather restricted definition of an experiment. The usage in the monograph considers a true experiment as involving at least two treatments - perhaps one active treatment and a control treatment; or two active treatments; together with randomized assignment of treatments to experimental units. This definition is conventional in the statistical literature but quite different from common administrative or bureaucratic usage where "experiment" may mean simply a try-out, a preliminary version of a program that will later be conducted on a larger scale. The usage adopted in the SSRC monograph is much narrower, and conforms to the usual statistical usage.

Two remarks about the design of social experiments may be apropos. One may well ask in respect to variations in the intensity of the treatment whether it might be prudent adding an extremely intensive, even an implausibly intensive treatment, to a design just in order to test whether any treatment at all of the character proposed, at any intensity of application would be effective. Since many social interventions seem weak in comparison with spontaneous counter-forces, it seems worthwhile to inquire whether it is the character of the treatment or merely its intensity of application that produces null or negative results. This is an argument for including treatments that would not be programatically feasible on a national scale since feasibility is of no consequence if even an unfeasibly strong treatment is shown to be ineffective. Secondly, one must face the question of representativeness in the choice of experimental units. It is important to ask: "representativeness for what"? Is the experiment being done for purposes of parameter estimation and generalization to some population? Or is the experimenter simply looking for treatment effects? The answer will determine whether representative sampling of subjects is important.

With the treatment program and the experimental design in hand, an organization must be developed to administer the treatment. For instance, in the food supplementation experiments it is necessary to provide transportation of the raw materials to the experimental site, a place to cook the supplement, a feeding center to which eligible children and mothers come every day, a staff to take care of preparation, serving and clean-up afterward and other matters. This is a rather different task from the technical problems of design or the strategic problems of treatment development and requires different talents. These may not be highly valued by academic institutions, but they are non-negligible. Careless treatment administration can invalidate an experiment just as easily as poor data collection.

Even before the experiment proper is off and running, one must give attention to the analysis and feedback of data for purposes of program revision, policy planning, or perhaps installation of a revised version of the program. This stage brings the cycle of program planning, test, revision and installation to a close.

One might think that, with such a restricted definition of what constitutes an experiment the list of social experiments would be very short. That is, one might suspect that there have been very few randomized, controlled experiments in which the treatments were genuine interventions into societal processes and not merely laboratory exercises. Not so. Robert Boruch's efforts on behalf of the SSCR project have turned up over 120 true randomized social experiments conducted within the last 20 years. This list will be included as an annotated bibliography in the monograph. It covers experiments on such topics as: social rehabilitation programs for juvenile and criminal offenders; law-related programs and procedures; rehabilitation programs in mental health; sociomedical problems and fertility control; assessment of educational and training programs; and many others. The contents of this bibliography demonstrate that randomized experimental tests of social interventions are feasible in a variety of program settings.

Many statisticians are interested in natural approximations to designed social experiments and one section of the SSRC monograph covers these. Donald Campbell, one of the authors of the monograph, has used the phrase "quasi-experiments" to characterize certain natural situations in which something close to an experiment spontaneously and unintentionally occurs. Often, some legislative or administrative change is the virtual equivalent of a deliberately imposed treatment that affects a large segment of a population all at once. One can then observe a time series being interrupted by the change, and look to see whether certain effects have occurred - i.e. whether there have been changes in some dependent variable. An illustration is the administrative decision to require breathalyzer tests of motorists in Great Britain after a certain date, which can be analyzed in relation to rate of car accidents and highway fatalities.

Most of the examples in the SSRC monograph are drawn from the last decade or so of social interventions, but quasi-experiments are, of course, not unique to our own time. I recently discovered a much earlier example that may be of particular interest to this audience because it involves the well known early English bio-statistician, William Farr. In the mid 19th century metropolitan London was the scene of a number of recurrent epidemics of cholera. At that time the disease was not well understood and the mode of its transmission (through water contaminated with fecal matter from infected persons) had not been discovered. Farr had studied cholera epidemics and developed a very

interesting theory about the mode of transmission of the disease*. Farr's study of the data from the 1849 cholera epidemic led him to the conclusion that cholera incidence was related to micro-differences in elevation in various parts of London mediated by differences in miasmas, or atmospheric factors that varied with altitude. His analysis suggested a neat linear progression in incidence varying inversely with 20 foot differences in altitude in the city, leading him to conclude that the lower the altitude the denser the miasmas and hence the higher incidence of cholera. In 1852 the metropolitan government of London passed a law which required that all river water supplied by the private water companies for domestic use must be drawn from the Thames above Teddington Lock, or from tributaries of the Thames above tidal influence. At the time of the 1854 cholera epidemic only one company, the Lambeth Waterworks had complied with the new regulation and had, thereby, suddenly changed its source from one of the most to one of the least contaminated by sewerage. The further important fact is that in a number of districts of the city the Lambeth Company competed directly, street by street, house by house, with the Vauxhall Company which had continued to draw its water supply from a highly polluted portion of the river. In other respects, the social and sanitary conditions of the patrons of the two companies were virtually identical. This situation constituted a vast quasi-experiment involving nearly half a million people for whom the source of drinking water was the most important difference among all social and environmental conditions. By comparing cholera incidence among the customers of the Lambeth Waterworks in the 1854 epidemic with the same customers in the 1849 epidemic; and comparing the experience of patrons of the two companies during the 1854 epidemic, Farr was able to identify the role that drinking water played in the transmission of cholera. (It is of perhaps incidental interest to note that the outcome of this quasi-experiment did not persuade Farr to withdraw his miasmal theory. He simply expanded its scope to include water as well as air among the malignant miasmas).

Quasi-experiments present certain interesting problems of statistical analysis, which arise mainly from the non-random assignment of treatments to units, but that is not the only reason for preferring true randomized experimental design. Appropriate "natural experiments" cannot be counted on to appear in timely fashion to help shape social interventions, programs and policies. Administrators and planners may need to have recourse to randomized experiments for a variety of purposes. One purpose is to estimate parameter values, as in the New Jersey Negative Income Tax Experiment where the problem was to estimate the size of work disincentive effects of a non-conditional income grant. Another purpose is to compare two or more treatments, e.g. the effects on mental development of protein supplementary feeding alone, with the effects obtained from supplying intellectual stimulation along with the protein. A third purpose is to test a concept or claim, as, for example, in the "performance contracting" experiment where certain commercially developed instructional programs in reading and arithmetic were experimentally compared to traditional public school methods to test the claimed superiority of the former.

All of these are, in a sense, specialized and subsidiary purposes that can be encompassed as special cases of the kind of experimental program development, test and revision that was sketched out above and proposed as a substitute for simply installing a program and then evaluating it after the fact. Given the history of social intervention with its abundance of uncertain and sometimes, unintended outcomes, it would seem prudent to learn these lessons on a small, experimental level before going into a nation-wide (or state-wide or company-wide) program that is almost guarenteed to have some flaws, and to be difficult to change or to withdraw.

* For further information about Farr, his theory, and the London cholera epidemics, the reader may consult: Eyler, JM: William Farr on the cholera: The samitarian's disease theory and the statistician's method. Journal of the History of Medicine and Allied Sciences 28:79-100, April, 1973.

DISCUSSION

Oscar Kempthorne, Iowa State University

I am sure that all of us appreciate the importance of the activity under discussion, that is, social experimentation. I hope that the complexities of the total process of defining an experiment, performing it, collection of appropriate data, data interpretation, and, finally, the drawing of conclusions, are adequately appreciated. I mention this specifically because I am led by my reading both in the technical literature and in the semi-popular and popular press to the opinion that the complexities are not appreciated.

The past two decades of the United States and of the world have been remarkable for the wide concern that has been felt for human problems. It is interesting in this connection to note that concern about poverty "of the masses" and all the difficulties of people all over the world was felt before this century and even before these two decades only by very few. Indeed, it is more than interesting, it is quite remarkable in the history of mankind. One hundred years ago, the few people who were concerned were considered by most of "educated" humanity to be crackpots. There were thousands of so-called educated people, e.g. graduates of universities, who simply ignored the problems, and of those who did, the majority approached the problems with a value system which from our present point of view we can only term appalling. One could write a fascinating book on this. An example which struck me a few years ago was Galton who was a near-genius to be sure, but whose writings reflect strongly the prejudices and value system of academia of Great Britain of the 19th century.

Several of the crackpots performed what were and are called experiments, though, in fact, they were not experiments in the modern sense of comparative experiments, but consisted almost entirely of the implementation in a community of the ideas of a leader, with no defined "treatment' protocol, no replication and no control in the sense that the idea is used in the modern comparative experiment. The books on these "experiments" make fascinating reading to be sure, but there is almost a complete absence of any sort of experimental inference.

The area of social experimentation is of vast importance because every Dick, Tom and Harry is concerned with social programs. Legislators promulgate new social programs by the score. Our governmental apparatuses are involved deeply and politics is all-pervasive. The intrusion of politicians and the bureaucracy is inevitable. There is considerable risk that political influence will be exerted to induce conclusions from social experiments that are acceptable from a particular political viewpoint. The piper may try to call the tune, the piper in this case being a political high appointment in the supporting federal agency.

My contact with social experimentation at a professional level has been almost zero, and the only basis for my contributing to the present session is that I have had intimate contact with experimentation in biology and agriculture and moderate contact with physical science and engineering. I am able, therefore, only to talk about general principles of experimentation with the hope that some of the remarks I make will have some relevance.

I was involved many years ago in what may be thought of as a very simple social experiment. The question was what would be achieved by supplementing the diet of children in grade school by giving them a glass of orange juice each morning at school. This is ludicrously simple in comparison with the experiments that are presented in this session. But I found great difficulty in developing a rationale for the design, for the choice of data to be taken, and for the analysis of the data. I mention this because the problem is extremely simple relative to the type of problem considered as social experimentation. The difficulties are compounded by the eagerness of workers, the public in general, politicians and public executives for definitive answers. I was struck in this connection by a statement in the Jensen paper. This paper has aroused a huge amount of controversy, and, I believe, justifiably so, for a multitude of reasons. In connection with the present discussion, we find the strongly assertive statement:

> "Compensatory education has been tried and it apparently has failed."

This is followed by a long exposition of a hypothesis for the failure. I have to record my antagonism to this kind of statement in science generally and to this particular statement in its specific context. I ask:

> Is compensatory education a well-defined treatment, like for instance a treatment protocol in an acceptable medical experiment?
> Is it not the case merely that some very ill-defined procedures which have some of the appearances of what we all think of as compensatory education were tried in an uncontrolled experiment?
> What were the possible biases of the studies?
> What was the sensitivity of the studies?

I am of the opinion that the making of such a statement is more in the nature of demagoguery than reasoned scientific evaluation. I question also whether it is consonant with the social responsibilities of scientists. Quite apart from any other aspect, it is clear to the near-idiot that compensatory education does succeed to a considerable extent. We can see this at all levels. Reading recently C. P. Snow writing on G. H. Hardy, one of the top mathematicians of the world of this century, I saw that the acquisition of the

best possible tutor was an important act in the seeking of high status in the Cambridge University Mathematical Tripos. Why do we seek teachers to give compensatory (and improving) education to our children - we of the "upward bound" or "upward oriented" sections of society? Because we know that compensatory education of some sort does in fact work. So what is the status of the assertion of Jensen? If there is a germ of truth in it, as there may well be, let that germ be stated clearly.

It is essential that science be responsible and that it state knowledge and lack of knowledge very precisely, particularly in the context of deeply human affairs. The same holds with regard to heredity and environment, but this is outside the present arena.

In thinking about experimentation, it is useful, I believe, to run over the spectrum of experiments from Galileo's trials with balls rolling down inclined planes, to rates of reaction in chemical kinetics with varying concentrations of reactants, to determining the effects of nutrients on plants, to determining the effects of nutrients on animals, to determining the effect of physical nutrients on man, to determining the effects of physical, mental and economic nutrients on humans, which is what the present discussion is aimed at. This whole spectrum may be partitioned in various ways, but one partition is, I think, of critical importance. On one pole we have the experiments on physical objects, or on objects which do not have mental apparatus. On the other pole, we have objects or entities which cannot be observed, let alone experimented on, without the observation process itself producing an effect. The phenomenon was surely known back to antiquity in the case of humans and animals. It is interesting that it came to light so recently in physical science with the Heisenberg Uncertainty Principle.

If experimentation is done on humans, they know that they are being studied and subjected to chosen stimuli. This, alone, regardless of the nature of the stimuli, will affect their behavior and their reactions. In the case of humans and drugs, it is often possible to use placebos to measure some of the effects irrelevant to the drug being tested. But in the areas of social experimentation that are now under discussion, it seems very difficult to obtain any indication of the effects of experimentation, regardless of the stimuli. It will be not at all unusual for an experiment to show effects that are not found in a general societal program, because of such effects, which have been given, I surmise, special names. Social experiments are in a real sense psychological experiments with all the difficulties that these encounter.

Another factor is the occurrence of very long term effects. The simple physical (e.g. temperature varying) experiment on a piece of material, or the simple experiment on the effect of plant nutrients on the yield of, say, wheat or potatoes gives a nearly complete answer in a short interval of time, e.g. a season. But in the case of social experimentation, the experimental material may show effects for decades. The immediate effects may be encouraging, but these may 'wash out' over time, or there may be effects over a long period of time that are evinced only over time and cannot be predicted from the short term experiment. I feel that general remarks of this sort are necessary, because it is so easy to say "Let's do an experiment and we will see what happens."

It is also relevant to note that even in areas which do not present the difficulties alluded to, the progress of knowledge through experimentation has been very slow. I surmise that no one has made a census of the number of experiments that have been conducted on a "simple" problem, such as what nutrients to feed to wheat or to pigs. One would think that a simple experiment involving say, 50 pens of pigs would give all the necessary answers. But the fact is that no one such experiment or not even 100 such similar experiments have given definitive answers. So to hope that 2 or 3 social experiments will give definitive answers is very naive. This needs to be said because our citizenry and our politicians (and perhaps some of our scientists) will think so.

Related to this is the additional fact that the structure and technology of present-day experiments has been developed only after very laborious pains-taking steps. Technique of experimentation has taken a long time to develop and is still developing. So it may be surmised strongly that we do not know how to do social experiments. We learn how to do experiments only by doing them and learning from mistakes. So there is a very long road ahead for the social experimenters, and it is critical that our citizenry and politicians have some appreciation of the vast difficulties.

There were in the 20's and 30's several books on what was called field plot technique, i.e. how to perform agricultural and biological experiments. I would like to see a book on "social plot technique" if I may use a crude analogy, by which I mean how to choose experimental units, how to take care of border effects, how to measure the experimental entities, and so on. I say this because the technology and the body of now-accepted principles came about only with many workers with different ideas and working in different directions. Many of the experiments were aimed at the building of knowledge, and the formulation of action programs came later. In the social area there is now the natural desire to do one or two simple experiments and to hope to obtain from them action programs which will solve the ills of society. If furthermore, one or two simple experiments do not give a social prescription, then it is inevitable that many will say that experimentation has failed.

In view of my background of knowledge and experience, I have to confine myself to the philosophy, logic and technology of exper-

iments. The simple notions that arose in agricultural experiments have, I think some force in potential social experiments. The initial step is surely the making of a choice of what is termed experimental units, and the related choice of a pattern of imposition of treatments on the experimental units. In the case of an agronomic experiment, the experimental units are simply field plots. In the case of an animal experiment, the unit may be a single animal or a pen of animals. In social experiments, the unit could be a single person, a family, the families on a city block, the families of a certain type, of a town of a certain type, and so on. If it is possible to give one person one treatment and another person another treatment, the person could be the experimental unit. In contrast, it may be that one can apply a treatment only to the family, or the household, or the apartment block, or the city. The next step is to realize that it is impossible to obtain units which are identical. There will be variability between units. The prescription that has been accepted universally (apart from some of the neo-Bayesians) is that the only valid way to obtain some statistical control of this variability in comparisons of treatment effects is to adopt the two ideas of blocking or stratification of the units. So we have designs like the randomized block design, the split-plot design, the Latin square design, the incomplete block design. The logic of the process was given in highly heuristic terms in the classic by R.A. Fisher, "The Design of Experiments". The simple point is that the experiment is regarded as a single trial of a population of trials, in which the contribution of variability between units to comparisons between treatments is guaranteed by the conduct of the experiment. This is to be contrasted with the control of variability between experimental units and its effects by means of assumed linear models and processes of linear or non-linear least squares or some other method of fitting a model to data. The thrust here then, is to obtain validity of the experiment, in the sense that those forces which contribute to treatment differences must contribute equally to the estimation of error by which one assesses the reality of observed treatment differences.

It is important in this connection to note that if an experimental unit consists of several, say 5 animals, then the variability of response within the experimental units may be strongly misleading as a measure of the variability to which treatment differences are subject. If then a school were to be the experimental unit, with different treatments on different schools, the variability between observed treatment differences would depend on the variability between schools, over and above the variability between pupils within schools. This type of thinking, if it be accepted, has a devastating effect on one's ideas of size of experiment. Two social programs that are compared on 4 cities, with 2 cities receiving one program and 2 cities receiving the other, has only 2 replicates with regard to the variability between cities. And this is the case whether the cities have one thousand or one million inhabitants. If, however, one can assume that there are no city differences but only variability within cities, then the replication with regard to that source of error will be the relevant one and may well be large.

It is also important to make a distinction between exploratory and confirmatory experiments. In the exploratory experiment one will impose several different treatments, and one will subject the resulting data to as wide a variety of analyses as one considers to be worth exploring. In the confirmatory experiment on the other hand, the whole protocol of the experiment both in design and in analysis must be prespecified. In the noisy sciences almost every experiment has two aspects, in that insofar as it is used for confirmation the analysis must be prespecified, but insofar as it is exploratory, the field of possible analyses is wide open.

If an experiment-experimenter interaction produces a new idea, then that new idea becomes an input for a new confirmatory experiment. The philosophy of knowledge and statistical theory have not achieved, it appears, a mode by which hypotheses suggested by a set of data may be confirmed on the basis of that set of data. On the other hand, significance tests may be applied which enable one to make a judgment of the extent to which an apparently aberrant result is actually aberrant on the basis of an assumed model.

So I see a wide variety of social experiments, many of which will be inconclusive, most of which I hope, will be suggestive. I see a great need for repetition of experiments which have been suggestive. I see also a need for the formulation of the ethics of social experimentation, just as there has been a formulation of ethics for medical experimentation. I surmise that there are very considerable difficulties in this respect which cannot be ignored. The notion of an era of social experimenters using our citizenry as guinea pigs is offensive, but the idea that experimentation should not be done will delay the development of the sort of society that we seek.

USE AND EVALUATION OF SYNTHETIC ESTIMATES

Maria Elena Gonzalez, U.S. Bureau of the Census

Definition of synthetic estimates

An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas, on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates. For the smaller areas, the estimates are no longer unbiased. However, it is possible to measure an average mean square error (MSE) for this set of estimates.

The simplest synthetic estimates are obtained by assuming that for the statistic of interest the mean value in the large area applies to each subarea directly; more refined estimates can be obtained by making this assumption for subgroups of the population. In the case when subgroups of the total are used, they should be nonoverlapping and exhaustive; the statistical estimates for the subgroups of the larger area are combined using independently known weights for the smaller area (e.g., as found at the time of the census) to obtain synthetic estimates for the smaller areas.

One is interested in estimating a characteristic, X. Identify j subgroups in the population, which are nonoverlapping and exhaustive. From the larger area we obtain estimates, $x_{.j}$, for $j=1, 2, \ldots, G$.

A synthetic estimate is desired for subarea i, which is within the larger area. From the latest census we have weights p_{ij}, such that

$$\sum_{j=1}^{G} p_{ij} = 1.$$

The synthetic estimate, x_i^*, for characteristic X and subarea i is defined as

$$x_i^* = \sum_{j=1}^{G} p_{ij} x_{.j} \qquad 1)$$

This estimate associates the characteristic $x_{.j}$ of the larger area with each of the subareas i.

Use of synthetic estimates

Synthetic estimates are used primarily to develop small-area estimates when sample sizes are too small to give reliable results directly. Some examples of recent uses follow.

1) The National Center for Health Statistics has developed synthetic State estimates of disability based on the Health Interview Survey data. National rates of disability for 78 subgroups defined in terms of age, sex, size of household, income, industry, etc., were obtained from the data collected in the Health Interview Survey. These disability rates were weighted by the corresponding population in individual States, from the 1960 Census of Population, to derive synthetic State estimates of disability. [1]

2) The Bureau of the Census has used synthetic estimates for the imputation of population for units reported as vacant in the 1970 Census of Population and Housing, but which were actually occupied. A subsample of the housing units reported as vacant in the 1970 Census of Population and Housing was selected and interviewers were sent to these units to determine how accurately that determination had been made. About 11 percent of the housing units reported as vacant were determined to have been occupied at the time of the census. Separate estimates of such error rates were prepared for twelve geographic areas within the United States. Within each area the rate was applied to each enumeration district in the census and the applicable percentage of vacant units was converted to occupied units. The estimates of the error rates for areas such as cities, counties or States were synthetic estimates. [2]

3) In order to study the properties of synthetic estimates, an experiment was conducted to develop unbiased and synthetic estimates of unemployment for SMSA's for monthly, quarterly and annual estimates based on the Current Population Survey (CPS) data. A comparison of the reliability of the two types of estimates revealed that for monthly data the synthetic estimates were preferable, while for annual data the unbiased estimates were preferable; for the quarterly data, the two were of about equal reliability. [2]

4) In the 1960 Census of Housing enumerators were instructed to rate the physical condition of each housing unit into one of three categories: "sound," "deteriorating," or "dilapidated." One important purpose of this was to provide data on substandard housing defined by Federal and local housing agencies as comprising units lacking complete plumbing facilities plus units which were dilapidated but had all plumbing facilities.

In the 1970 Census of Housing information was again obtained about plumbing, but synthetic methods were used to develop estimates of housing units which were dilapidated with all plumbing facilities (DWAPF). To obtain these estimates census data on housing units with all plumbing facilities were multiplied by estimated proportions of dilapidated housing units which had all plumbing facilities, as derived from a post-census survey, Components of Inventory Change (CINCH). 1/ From CINCH, estimates of DWAPF housing units were obtained for specified subgroups for 15 selected large SMSA's and for four balance of regions of the U.S. Synthetic estimates for the smaller areas within these nineteen geographic areas were derived using the corresponding set of DWAPF proportions.

Evaluation of synthetic estimates

Synthetic estimates are biased; to evaluate their reliability one can use the MSE, which can be expressed as the sum of the variance and the

square of the bias:

$$MSE(x_i^*) = \sum_{j=1}^{G} p_{ij}^2 \sigma_{x.j}^2 + (X_i^* - X_i)^2 \qquad 2)$$

where

$\sigma_{x.j}^2$ is the sampling variance of estimate $x_{.j}$,

X_i is the "true value" of the statistic for subarea i, and

X_i^* is the expected value of the synthetic estimate for subarea i.

The estimate given in formula 2 assumes that:

a. the p_{ij}'s are fixed and measured without error; and

b. the cov $(x_{.j}, x_{.k}) = 0$, for $j \neq k$.

In general, the values of X_i are not known and consequently the MSE of an individual synthetic estimate cannot be calculated for a particular area "i." However, if we establish M subareas within the survey population, the average MSE of the synthetic estimate over the M subareas (which may be unequal in size) can be estimated from the sample. Let

$$E\left[\frac{1}{M} \sum_{i=1}^{M} (x_i^* - X_i)^2\right] = \alpha \qquad 3)$$

The average MSE can be estimated by using the following approximation:

$$\hat{\alpha} \doteq \frac{1}{M} \sum_{i=1}^{M} (x_i^* - \sum_{j=1}^{G} p_{ij} x_{ij})^2 - \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{G} p_{ij}^2 (1-2f_{ij}) \sigma_{x_{ij}}^2 \quad [3] \qquad 4)$$

where

x_{ij} is the estimated statistic from the sample for subclass j and area i,

f_{ij} is the sample estimate of the proportion of the total for the j-th subclass that is in the i-th subarea, that is

$f_{ij} = n_{ij} / \sum_{i=1}^{M} n_{ij}$, and

$\sigma_{x_{ij}}^2$ is the sampling variance of estimate x_{ij}.

The interpretation of the square root of the mean square error is not altogether clear. Probability statements which can be made using the standard error for an unbiased estimate do not necessarily hold when the estimates are biased and the root mean square error is used as a measure of reliability. To try to understand the situation, an empirical study of the root mean square errors was made, using 1960 housing data. Synthetic estimates for a census were compared with actual measurements of the item for the same census to obtain an estimate of the bias. For various groupings of areas we then computed an average root mean square error (RMSE); this estimate together with the distribution of the biases was then used to compare an empirical distribution of the biases of synthetic estimates with the normal distribution.

As part of the publication of the 1970 Census of Housing, data on housing units dilapidated with all plumbing facilities collected in the 1960 Census of Housing were available for comparison with a set of synthetic housing estimates of DWAPF derived from the same data. The average mean square error for a set of M areas is given by

$$\text{Average MSE} = \frac{1}{M} \sum_{i=1}^{M} (x_i - x_i^*)^2 \qquad 5)$$

where

x_i is the census estimate for area i.

The use of formula 5 to estimate the average MSE assumes that the second term of formula 4 is negligible. This assumption is reasonable for large areas. The square root of the average MSE gives the estimate for the RMSE.

For all States we have two estimates of dilapidated housing units with all plumbing facilities in 1960; the 25-percent census estimate and a synthetic estimate based on a particular set of subgroups. The difference between these two estimates will be used as an estimate of the bias of the synthetic estimation procedure. Table 1 shows estimates of the proportion of the set of synthetic estimates for States with a relative bias within specified values. The relative bias for an area is defined as the difference between the synthetic estimate and the census estimate divided by the synthetic estimate.2/

Table 1. Distribution of Relative Biases of Synthetic Estimates for States

State estimate	Number of areas	Proportion with relative biases				
		0-9%	10-19%	20-29%	30-49%	50%+
1,000-2,499	7	0.14	0.29	0.29	0.14	0.14
2,500-4,999	6	0.50	0.17	0.17	0.17	0.0
5,000-9,999	13	0.23	0.46	0.15	0.08	0.08
10,000-19,999	16	0.38	0.38	0.19	0.06	0.0
20,000 or more	8	0.38	0.38	0.25	0.0	0.0
Total	50	0.32	0.36	0.20	0.08	0.04

This table shows that the proportion of estimates with large relative biases diminishes as the size of the synthetic estimate increases. For example, for synthetic estimates between 1,000 and 2,499 DWAPF housing units, 57 percent have relative biases of at least 20 percent; however, for synthetic estimates of over 10,000 units only about 25 percent have relative biases greater than 20 percent. When we consider the State synthetic estimates for all States, we note that 32 percent have relative biases of 20 percent or more, 12 percent have relative biases of 30 percent or more and 4 percent have relative biases of 50 percent or more. The average number of DWAPF housing units for States is about 13,000; the estimated average root mean square error is about 2,500; the ratio of the $\overline{RMSE}$ divided by the average size of State synthetic estimates of DWAPF housing units is 0.19. A high variability of the synthetic estimate is shown by the fact that the $\overline{RMSE}$ divided by the mean is about 20 percent. This shows that the synthetic estimates obtained do not account for a large part of the variability among areas. The synthetic estimates of housing units DWAPF are computed using a particular set of subgroups, defined in terms of tenure, race of head of household and other characteristics related to the quality of housing unit. The use of other subgroups would produce a different set of synthetic estimates.

From the point of view of ascertaining whether the average root mean square error can be used to make probability statements the results are more encouraging. Table 2 gives some comparisons of the distribution of the difference between State synthetic estimates of DWAPF housing units for 1960 and the estimates reported in the 1960 census. The first two columns of the table show the expected percentage of the normal distribution at different multiples of the standard error (σ). For example, 95 percent of the normal distribution is expected within two standard errors of the mean. The empirical distributions of the biases of the synthetic State estimates of DWAPF housing units are given in columns 3, 4 and 5. For example, 48 percent of the biases for estimates of total for States are less than one-half the estimated $\overline{RMSE}$.

Table 2. Comparison of Empirical Distribution of the Biases of State Synthetic Estimates of Dilapidated Housing Units with All Plumbing Facilities with the Theoretical Normal Distribution

Multiple of standard error (σ)	Normal probability	Distribution of bias of state estimates (n = 50)		
		Total Aver = 12,970 $\overline{RMSE}$ = 2,490	Inside SMSA's Aver = 8,170 $\overline{RMSE}$ = 1,540	Outside SMSA's Aver = 4,800 $\overline{RMSE}$ = 1,340
0.50	38%	48%	50%	62%
0.75	55	62	62	68
1.00	68	74	68	74
1.25	79	86	82	84
1.50	87	88	90	88
1.75	92	88	90	88
2.00	95	92	94	94
2.25	97.6	94	94	98
2.50	98.8	96	96	98
3.00	99.7	100	100	98

The empirical distributions of the biases of synthetic State estimates are closer to the mean (on the average) for values within one standard error of the mean, than expected for the normal distribution. For example, for half a standard error the normal distribution expects to cover about 38 percent of the distribution; for State totals, the empirical distribution actually includes 48 percent of the distribution; for estimated units within SMSA's, the empirical distribution includes 50 percent of the distribution and for units outside SMSA's the empirical distribution includes 62 percent of the distribution. However, for values which are more than two standard errors from the mean, the empirical results are reversed: the frequency of synthetic estimates with biases more than two standard errors from the mean is greater than expected for normal distributions; for State synthetic estimates about 8 percent had biases which differed by more than two standard errors from the mean. That is, on the average there are more outliers for synthetic estimates than would be expected for a normal distribution.

Table 3 shows empirical distributions of the biases of the estimates of DWAPF housing units for non-Negro renters in counties within SMSA's. The distribution was computed separately depending on the magnitude of the estimate of DWAPF housing: less than 100, 100 to 499, and 500 or more units. It is possible to carry out the analysis separately for the three groups through the use of formula 5 for the average root mean square error. The results reveal a similar

pattern to the results for State estimates given in Table 2. For values within one standard error the empirical distribution gives conservative estimates of the probability of occurrence, except for the distribution of DWAPF housing units with 500 or more units. However, for values at three standard errors we find more outliers than expected for the normal distribution. For example, for synthetic DWAPF estimates less than 100, the normal distribution expects only 0.3 percent of the cases to be further than three standard errors, but we find that 2.5 percent of the values have biases larger than three times the average root mean square error.

Table 3. Comparison of Empirical Distribution of the Biases of Synthetic Estimates for Non-Negro Renters in Counties within SMSA's of Dilapidated Housing Units with All Plumbing Facilities with the Theoretical Normal Distribution

Multiple of standard error (σ)	Normal probability	Distribution of bias for non-Negro renters in counties within SMSA's		
		DWAPF <100 $\overline{RMSE}$ = 39 n = 160	DWAPF (100-499) $\overline{RMSE}$ = 125 n = 219	DWAPF (500+) $\overline{RMSE}$ = 300 n = 84
0.50	38%	50%	49%	37%
0.75	55	66	66	52
1.00	68	84	79	64
1.25	79	89	85	75
1.50	87	93	89	85
1.75	92	94.3	92.2	89.2
2.00	95	95.6	94.5	95.2
2.25	97.6	96.2	96.8	98.8
2.50	98.8	96.2	97.7	98.8
3.00	99.7	97.5	98.6	98.8

Conclusions

Census data allow us to compute synthetic estimates and to compare them directly to the census estimates. Therefore, the biases of synthetic estimates can be obtained and their distribution analyzed directly.

The results presented comparing 1960 estimates of dilapidated housing units with all plumbing facilities with synthetically derived estimates show that the synthetic estimates are highly variable, but that the distribution of their biases is not too far from normal.

The analysis presented is based on a particular set of synthetic estimates; alternative sets using other variables should be investigated in order to be able to select the subgroups which account for a large proportion of the variability of the local area estimates, with an aim toward improving local area estimates. The results presented here, based on a particular set of synthetic estimates, may not necessarily generalize to possible alternative sets of synthetic estimates.

FOOTNOTES

1/ The change in procedure in estimating DWAPF housing units was necessary because a majority of housing units in the 1970 Census of Housing were enumerated by a mail-out, mail-back procedure; in addition, studies of these data for 1960 indicated that statistics based on enumerator ratings are highly unreliable.

2/ The synthetic estimate was used as denominator, instead of the reported estimate, because in 1970 the synthetic estimate was the only one available.

REFERENCES

[1] "Synthetic Estimates of Disability," published in 1968 by the National Center for Health Statistics, PHS publication No. 1759.

[2] A more detailed discussion of this project is available in "Estimation of the Error of Synthetic Estimates," by Maria Elena Gonzalez and Joseph Waksberg, presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, August 18-25, 1973.

[3] Appendix 1 of "Estimation of the Error of Synthetic Estimates," by Maria Elena Gonzalez and Joseph Waksberg gives the derivation of the approximation given in formula 4.

RECENT DEVELOPMENTS IN ESTIMATION FOR LOCAL AREAS

Eugene P. Ericksen, Institute for Survey Research, Temple University

1. Introduction

The regression-sample data method of postcensal estimation is a procedure by which one can combine sample survey data with symptomatic information to obtain local estimates of the criterion variable being measured by the survey data. This method has been tested extensively using population growth as the criterion (Ericksen, 1973a, 1973b), first, for the period beginning in 1960 and ending in 1964-67, and then more extensively for 1960 through 1970. The steps of the procedure in the latter test were as follows:

a. Sample estimates of population growth were obtained for the primary sampling units selected into the national sample of the Current Population Survey. These 1970 estimates of current population were divided by the corresponding 1960 Census populations giving sample estimates of 1960-70 population growth.

b. Symptomatic indicators, in this case 1970/1960 ratios of births, deaths, and school enrollment, were compiled for the sample psus and a multiple regression equation was computed using the sample estimates of population growth as the dependent variable. A second equation was then computed using the series of ratio-correlation estimates calculated at the Population Division of the Bureau of the Census as a fourth symptomatic indicator.

c. Values of the symptomatic indicators for counties were substituted into the regression equations and estimates were made of the 1960-70 population growth. This step was carried out for 2,586 counties in 42 states.

Corresponding estimates for these counties were made at the Population Division using four standard demographic techniques which have traditionally been used to estimate population growth. Of these techniques, the ratio-correlation technique was the most accurate. Little was gained from averaging estimates of two or more standard techniques. The regression estimates produced by our combination of sample data and symptomatic information were more accurate than those of any single or combination of standard techniques. This was particularly true when the series of ratio-correlation estimates were added as a fourth symptomatic indicator. There were moderate reductions in the mean error, but the greatest gain was in the reduction of the number of large errors, which was over 20 per cent. These results are presented in Table 1.

Some of the prominent features of our new method of postcensal estimation are the following:

a. Estimates of population growth have been shown to be more accurate. Part of the reason for this gain is that it is not necessary to make any assumptions concerning the nature of relationships beyond those of least squares linear regression. One of the difficulties of the ratio-correlation technique, for example, is the assumption of the continuance of past relationships.

b. Other series of estimates can be incorporated as symptomatic indicators. By including the series of ratio-correlation estimates as a symptomatic indicator, we have a way of correcting for the bias of ratio-correlation which arises from assuming the continuance of past relationships.

c. There is a procedure by which the mean squared error of the regression estimates can be calculated. Given this facility for measuring error, we can systematically test various combinations of symptomatic indicators to determine the composition of the optimal set. Because of the presence of the within-psu sampling error, this does not necessarily include all available symptomatic indicators.

2. Current Activity at the Bureau of the Census

A project is currently under way at the Bureau of the Census to compute yearly estimates of population growth since 1970 by our regression-sample data method. A determined effort is being made to obtain symptomatic data for all counties in each of the 50 states and the District of Columbia. It now appears that births, deaths, and school enrollment will be available, but with some time lag, for counties in all but a small handful of states. Additional data on automobile registrations will be available for some states and it is also expected that data on income tax exemptions will be available for all states (Zitter and Word, 1973). Substantial gains are anticipated from the use of tax records, even outside our regression-sample data format. In view of the changing relationships among variables, and the possibility that other symptomatic indicators will become available, the following instructions are pertinent to potential users of the regression-sample data technique:

a. Applications of and experimentation with the ratio-correlation technique have shown conclusively that relationships among a given set of variables can be expected to change over time. We have shown that the series of regression-sample data estimates were relatively accurate when computed over a given ten-year period for particular sets of three and four symptomatic variables. However, the accuracy of the regression-sample estimates relative to those of other techniques could change for a shorter estimation period beginning in 1970. It is also possible that the most accurate regression-sample data estimates would be computed with a different set of symptomatic indicators in this period. We

can test this possibility by inspecting the mean squared error of the regression estimates and the correlations of the various indicators with the sample estimates of population growth.

b. In the absence of correlations between the sampling error and the value of the symptomatic indicators, the estimated regression equation using the sample estimates of population growth as the dependent variable is an unbiased estimate of the regression equation which would be obtained using Census tabulations of population growth if they were available. However, the presence of the within-psu sampling error will lower the observed values of the correlation coefficients. Low observed values of the correlation coefficients do not necessarily mean that the errors of the regression estimates will be large.

c. There are some unsolved problems regarding the inference from a sample of psus to a universe of counties. The mean squared error of the regression estimates refers to the accuracy of estimates for psus, when the units of interest may be counties. To the extent that counties are different from psus, reductions in the mean squared error for psus may not improve the accuracy of estimates for counties. A second unresolved problem has to do with specification errors arising from the distribution of the within psu sampling errors. If the size and direction of these errors vary systematically with values of the symptomatic indicators, the assumptions of linear regression may not be met. We have found this to be a minor problem in our application that resulted in larger errors for units with extreme growth rates, but the problem could be more important in other applications.

3. The Mean Squared Error

We have shown elsewhere (Ericksen, 1973a, 1973b) that the mean squared error of the regression sample data estimates can be expressed by the formula:

$$\frac{E(Y - \hat{Y})'(Y - \hat{Y})}{n} = \frac{(n - p - 1)\sigma_u^2}{n} + \frac{(p + 1)\sigma_v^2}{n} \qquad (3.1)$$

where σ_u^2 = the between-psu variance unexplained by the indicators,

σ_v^2 = the within-psu variance,

n = the number of psus in the sample, and

p = the number of symptomatic indicators.

When n is large relative to p, the mean squared error is determined by (1) σ_u^2, which decreases when new symptomatic indicators are added, and (2) the within-psu component of error which increases when indicators are added. If there were no within-psu component of error, optimal results would be obtained by maximizing p, i.e., by utilizing all available symptomatic information. We have found in our applications, however, that the presence of within-psu sampling variability has often meant that the optimal set of symptomatic indicators did not include all that were available.

In the test of 2,586 counties, there were seven symptomatic indicators available: births, deaths, school enrollment, and the four standard estimates. As shown in Table 2, where the ratios of the 1970 to the 1960 Decennial Census populations were the dependent variable, gains in the accuracy of regression estimates for psus were obtained by increasing the number of symptomatic indicators from four to seven. However, in the more realistic application, when the within-psu component of error was present, the increase from four to seven indicators actually brought about an increase in the errors. The mean error of the 2,586 county estimates increased from 4.2 per cent to 4.7 per cent. A similar result was obtained when six variables, with 51 observations (one for each state and the District of Columbia) were available.

The fact that the optimal set of indicators included four variables was due to the nature of the structural relationships and the size of the within-psu variance. We have evidence that these change over time, as shown in Table 3. In particular, for the Current Population Survey sample, the within-psu variance increased. This is because the CPS sample was based on the 1960 Census, and that patterns of subsequent growth were uneven, leading to variation in the size of sample segments within psus. This trend leads us to expect that more symptomatic indicators should be used in shorter time periods. On the other hand, the relationships among the variables appear to become stronger as time passes. In spite of the increasing within-psu variability which dampens the observed correlations, these observed correlations grew larger from 1963 through 1967. In shorter periods, changes in population size, as well as in the symptomatic indicators, appear to be due more to random fluctuations. As time passes, changes in the variables are larger, and the relationships among these changes more systematic. This leads to the contrary expectation that the optimal set of indicators would be smaller for a shorter time period. To determine the optimal set of indicators, we must estimate the mean squared error in each estimating situation.

Because the true values of the criterion variable are unobserved, the mean squared error cannot be estimated directly. To obtain equation (3.1), we must first compute the mean of the squared differences between the regression estimates and the sample estimates for the sample psus and then subtract an allowance for the within-psu sampling error. The mean squared difference between the regression and sample estimates can be expressed by the formula:

$$\frac{E(Y_o - Y)'(Y_o - Y)}{n} = \frac{(n - p - 1)(\sigma_u^2 + \sigma_v^2)}{n} \qquad (3.2)$$

To obtain (3.1) we need to subtract the term $(n - 2p - 2)\sigma_v^2/n$. In order to obtain a good estimate of the mean squared error, we clearly need to have a good estimate of σ_v^2.

When we reported earlier results (Ericksen, 1973b), we did not feel that a good estimate of σ_v^2 was available. We had computed half-samples defined by the eight rotation groups of the CPS (U.S. Bureau of the Census, 1963) and had overestimated the mean error of the sample estimates for psus. This is because sample segments within the CPS sample had not been placed equally into rotation groups within individual psus. However, when the half-samples were formed on the basis of sample segments without regard to rotation group, a better estimate was obtained. The derivation of equation (3.2) depends on the values of (1) the sampling error and (2) the structural errors of regression, along with the sampling errors being unrelated to the symptomatic indicators. Our technique for estimating the mean squared error is particularly sensitive to these specification errors, as the following illustration shows.

The practical question we faced in the 1970 test was whether or not improvements in accuracy over that given by the ratio-correlation technique would be obtained by adding births, deaths, and school enrollment as symptomatic indicators in a regression equation. We found that the ratio-correlation estimates accounted for 92.7 per cent of the variance of the actual 1970/1960 ratios of population of the sample psus. Adding the three symptomatic indicators, the coefficient of determination, R^2, was increased to .951, a clear increase in the explained and reduction in the unexplained variance. However, the increase in the explained variance of the sample estimates of 1960-70 population growth obtained by adding the three symptomatic indicators to ratio-correlation was much smaller, from 41.7 per cent to 42.8 per cent. This was due to the presence of the within-psu error which is not reduced by adding symptomatic information. The observed variance of the distribution of sample estimates before regression was .0438. Using the series of ratio-correlation estimates as a single symptomatic indicator, the mean squared difference of the regression and sample estimates as expressed by equation (3.2) was .0255. This was reduced to .0250 when the number of symptomatic indicators was increased from one to four. Our estimate of the within-psu variance is σ_v^2 = .0253. Subtracting the allowance for this component of error, our final estimates of the mean squared error are .0004 where the ratio-correlation estimate is a single indicator and .0001 with four indicators. This is a very small difference considering the size of the within-psu variance and the mean squared difference between the regression and sample estimates. A small fluctuation could have seriously altered the observed results. When the number of symptomatic indicators was increased to seven, the coefficient of determination was R^2 = .432, the mean of the squared differences was .0249, and the final estimate of the mean squared error, .0002. These differences are so small that one may be on safer, although less scientific, grounds simply to observe that the increase in R^2 from .417 to .428 is large enough to produce a good reduction in error while guessing that the further increase to R^2 = .432 is not, given the increase in the number of symptomatic indicators.

Some of the difficulties in estimating σ_v^2 arise from the fact that the within-psu sampling error is positively correlated to the growth rate, and hence to the values of the symptomatic indicators. The correlation between the actual error of the CPS estimate and the estimated within-psu variance is +.45. This affects the estimate both of σ_v^2 and the way we obtain an estimate of equation (3.1) from equation (3.2). A second source of error is the correlation between the within-psu error and the growth rate, which is + .06. This introduces curvilinearity, since the sample estimates of the fastest growing areas tend to be too large and those of the slowest growing too small, thus biasing the estimation of regression coefficients. One result of this was that estimates of areas with extreme growth rates had larger errors. This particular problem is covered in the literature on econometrics where the usual solution is to apply a transformation. We have attempted several such solutions, but have yet to find a transformation which allows us to reduce the errors of the extreme cases without increasing the errors of the majority of cases which have moderate values.

One obvious procedure for reducing the mean squared error of the regression estimates is to reduce the within-psu variance. This could be done by improving the within-psu sample design, or, as we will attempt to do, by introducing more sample data. In our program at the Census Bureau, we plan to eventually request tabulations from other government surveys such as HIS and NCS. This will reduce σ_v^2 where the psus in the various surveys are the same and reduce the ratio $(p + 1)/n$ and therefore the within-psu component of error in equation (3.1) in cases where the psus are different.

This necessarily reduces the errors of the primary sampling units, but the effects on county estimates are uncertain. To illustrate this point, when the regression equation with three symptomatic indicators, births, deaths, and school enrollment, was recomputed using the 1970/1960 Census population ratios as the dependent variable, i.e., setting σ_v^2 equal to zero, the mean error of the psu estimates was 2.8 per cent. This compares to the mean error of 3.2 per cent when the CPS estimates were the dependent variable. The difference between 2.8 and

3.2 per cent was due to the within-psu error. However, when the two equations were used to make county estimates, the mean error was 4.4 per cent in both cases. The Census ratio equation, computed without the within-psu error, had done a better job of making psu estimates, but the transition from psus to counties had become more difficult. When the distribution of errors was broken down by size of the 1970 county population, it was found that use of the Decennial Census ratios in place of the CPS estimates had reduced the mean error for all categories of counties with population greater than 25,000, but that the mean error had increased among counties smaller than 25,000. Counties in this last category were the majority of all counties but were least similar to the CPS sample psus which usually consisted of combinations of counties picked with probabilities proportional to the size of the total population.

4. New Strategies and Plans

Given the limited gains obtained from reducing the within-psu component of error, and our lack of success in finding suitable transformations to reduce errors, the most promising approach to reducing our errors appears to be the introduction of new symptomatic information. One variable which has been shown to reduce errors is automobile registrations. Data were available in the 1970 test for 2,223 counties in 32 states. A five-variable regression equation, also including births, deaths, school enrollment, and the ratio-correlation estimate was computed and county estimates made. The mean error of these estimates was 3.8 per cent and 122 errors were greater than 10 per cent. The corresponding figures for this set of counties for the four variable regression equation omitting automobile registrations were 4.1 per cent with 148 large errors and, for the standard series of ratio-correlation estimates, 4.5 per cent with 220 large errors.

Another promising, but as yet untested, variable is the number of exemptions on income tax returns. Changes in address of persons listed on income tax forms are to be used to estimate net migration and when added to recorded natural increase, could give extremely accurate estimates of population growth. It is quite possible that these estimates would be sufficiently accurate in themselves so that little gain would be obtained by computing regression-sample data estimates. But it is more likely that some bias will be introduced because of the characteristics of persons not listed on tax forms or whose likelihood of being listed on a form varies at point of origin and destination. In such a case, this bias could be corrected by using the tax estimate as a symptomatic indicator in a regression equation possibly including other symptomatic indicators with sample data as the dependent variable.

Finally, we have made plans to attempt to estimate other variables such as racial composition, unemployment, and median family income. Although births and deaths are available by race in many counties, the chief barrier faced here is the lack of symptomatic information. Data on wages and work force appear to be available in metropolitan areas, but we are still searching for symptomatic data available on a national basis. If such data can be found, we can combine our symptomatic information and sample data with estimates which can be generated by other means. One such series would be the synthetic estimates being discussed in this session.

BIBLIOGRAPHY

Ericksen, Eugene P. "A Method for Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas," Demography, 10, 137-160, 1973a.

Ericksen, Eugene P. "A Regression Method for Estimating Population Changes of Local Areas," 1973b, manuscript submitted for publication.

U.S. Bureau of the Census, Technical Report No. 7, "The Current Population Survey--A Report on Methodology," 1963. U.S. Government Printing Office, Washington, D.C.

Zitter, Meyer, and David Word. "Use of Administrative Records for Small-Area Population Estimates," 1973, presented at the Annual Meeting of the Population Association of America.

ACKNOWLEDGEMENT

The research upon which this paper is based was carried out in cooperation with the Bureau of the Census, in particular with Benjamin J. Tepping of the Research Center for Measurement Methods and with the Population Division, who supplied all the symptomatic information used in computing the regression equations as well as special tabulations concerning the distributions of errors of the ratio-correlation and other standard estimates. Special thanks are due to Lori Kessler and Harris Miller without whose assistance at Temple University this research could not have been carried out. The findings, recommendations, and conclusions in this paper are the sole responsibility of the author and are not necessarily endorsed by the U.S. Government. The data in this paper are the result of tax-supported research and, as such, are not copyrightable. The data may be freely reprinted with the customary crediting of the source.

Table 1: Relative Accuracy of Standard and Regression-Sample Data Estimates of Population Growth, 1960 to 1970, for 2,586 Counties in the United States

Procedure	Mean Error[1]	Number of Counties With Error 10 Per Cent or Greater
Vital Rates	7.4	673
Component Method II	7.2	645
Composite	5.9	407
Ratio-Correlation	4.6	264
Component Method II, Composite, Ratio-Correlation, Averaged[2]	4.7	249
Regression-Sample Data, 3 Symptomatic Indicators[3]	4.4	220
Regression-Sample Data, 4 Symptomatic Indicators[3]	4.2	194

[1] All estimates were multiplied by an appropriate constant in order to sum to a separately estimated 42-state total.

[2] This was the most accurate combination of the four standard techniques.

[3] The 3-variable equation used was $\hat{Y} = .158 + .218X_1 + .142X_2 + .520X_3$.

The 4-variable equation used was $\hat{Y} = .058 - .097X_1 + .045X_2 + .214X_3 + .745X_4$.

Source of Standard Estimates and Estimate of 42-State Total: U.S. Bureau of the Census, Current Population Reports, Series P-26, No. 21, "Federal-State Cooperative Program for Local Population Estimates: Test Results - April 1, 1970," Washington, D.C.: Government Printing Office, 1973.

Table 2: Mean Errors Obtained With Various Sets of Symptomatic Indicators

Units of Estimates are 444 Primary Sampling Units.

Number of Symptomatic Indicators[2]	Mean Percentage Error, Psus[1]	
	Dependent Variable 1970 Census/1960 Census	Dependent Variable 1970 CPS/1960 Census
3	2.83	3.20
4	2.60	2.92
7	2.11	3.24

Units of Estimates are 50 States and District of Columbia.

Number of Symptomatic Indicators[3]	Mean Percentage Error, States[1]	
	Dependent Variable 1970 Census/1960 Census	Dependent Variable 1970 CPS/1960 Census
3	1.22	1.64
4	1.08	2.16
6	1.07	3.91

[1] Mean percentage error, comparing regression estimate with Census tabulation. "Dependent variable" is that used to compute regression equation.

[2] Set of three indicators included births, deaths, and school enrollment. The fourth indicator was the ratio-correlation estimate, and indicators five through seven were the composite, component method II, and vital rates estimates.

[3] Set of three indicators included births, school enrollment, and work force. The fourth indicator was deaths, and indicators five and six were automobile registrations and income tax returns.

Table 3: Values of Estimated Within-Psu Variance of Population Growth and Coefficients of Determination, 1963 through 1967

Year	Within-Psu Variance[1]	Coefficient of Determination (R^2)[2]
1963	.0253	.016
1964	.0378	.021
1965	.0383	.085
1966	.0458	.117
1967	.0473	.264

[1] Computed as squared difference between random half-samples defined by rotation group.

[2] Three symptomatic indicators were births, deaths, and school enrollment in each case.

DISCUSSION

Richard Royall, The Johns Hopkins University

Both of these papers are good examples of the process of developing estimators using conventional finite population sampling theory. We can pick out three important stages in this process:

1. Assuming observations on certain variables are available, scratch your head and write down an estimate which has some intuitive appeal.

2. Try to get a handle on bias and variance. (Having done this a few times and produced a few estimates, compare their mse's. Find one estimate is better than another under certain assumptions about population parameters.)

3. Get a real population and try out the estimates to see which works better under various realistic conditions.

After, or along with, these three basic steps comes the secondary problem of measuring the uncertainty in an estimate. This usually boils down to finding a nearly-unbiased estimate of an approximation to the variance or mse. Unfortunately, these variance estimates rarely have the "face validity" or obvious reasonableness of the original statistic. For example, the synthetic estimates are in a gross sense reasonable. They obviously won't give really precise estimates, but they will be, if not in the right ballpark, at least in the right city. The variance estimate, on the other hand, might not even be on the right planet -- a negative variance estimate might be reasonably described as "lost in space".

I would like to see a different approach used, and I think the problem at hand, estimation for small areas, is one in which this approach would yield different and better results than the conventional one, particularly with regard to providing estimates of mean square errors to use as measures of uncertainty. This approach would begin not with an estimate, but with an attempt to express the basic relationships among the relevant variables through a probabilistic model. The model would then be used to generate estimates, provide a framework for comparing estimates, and to provide estimates of standard errors. Often the conventional intuitive estimates are optimal or nearly so under a simple probability model, but sometimes the model suggests practical improvements, especially in the conventional measures of uncertainty. Varying the model can give valuable insight into the robustness of estimators. This general approach has been called "the prediction approach" because, when viewed in the context of (super-population) probability models, many finite population inference problems are mathematically equivalent to classical prediction problems. "The prediction approach" actually has many facets -- simple linear least-squares [4,5], esoteric fiducial [2], and full-blown Bayesian [1] prediction techniques are only some of those available.

What would be the results of applying the prediction approach (least-squares variety) to the present problem? Two important general results I would expect are:

1. New estimators and new variance estimators for the old ones.

2. New insight into relationships among estimates already proposed, and increased understanding of their strengths and weaknesses.

Specifically ... I don't know what results would be obtained. The work has not, to my knowledge, been done. But some relevant comments can be made.

The "ratio-correlation" method and the "regression-sample data" method aren't so much two different methods as two different estimates, each more or less appropriate under its own prediction model. Although the two models do employ slightly different functions of births, etc. as regressors, the most important differences between these two estimates come not from different assumptions concerning the relationships among the relevant variables, but from different assumptions about available data. The ratio-correlation method is not allowed to use the sample data, while the regression method employs only data from the sample and the most recent census, ignoring the previous census. In both models the total for a local area at one time is represented as a multiple of the total at an earlier time plus an error whose variance is proportional to the square of the earlier total. (We might ask whether a different error-variance might be more appropriate. If it is, this would suggest different estimates.) The multiplicative factor for a given area is a function of various bits of data concerning births, deaths, number of school children, etc. in that area. In this factor are certain coefficients which change over time. The "ratio-correlation method" uses estimates of out-of-date coefficients, while the "regression-sample-data" method uses less precise estimates of more timely coefficients.

When the "ratio-correlation" estimate is used as a "symptomatic indicator" in the "regression-sample data" estimate, we are, in effect, using a particular linear combination of estimates of the "old" coefficients and the "new". I think a formal model, in which coefficients for one time interval are stochastically related to those for an earlier interval, would be quite useful in evaluating this and other estimates based on all the data, from both censuses as well as the sample.

In much the same way, the choice between direct estimation and imputation in the synthetic estimation paper is really the choice between a high-variance estimate of a directly relevant parameter and a low-variance estimate of a different quantity. The choice need not be made -- surely a combination of the two is better than either taken alone. A probability model can express the

relationships whose existence makes the whole notion of "imputation" reasonable. Such a model would generate (via standard linear prediction techniques) statistics which would give proper weight to both direct and imputed estimates.

I think, however, that one of the possibilities suggested by Gonzales and Waksberg in their Vienna paper [3] is more promising -- before really good local area estimates are produced, the synthetic estimation approach must move towards Ericksen's in making greater use of available local area variables.

REFERENCES

[1] Ericson, W.A., "Subjective Bayesian Models in Sampling Finite Populations," Journal of the Royal Statistical Society, Ser. B, 31, No. 2 (1969), 195-224.

[2] Kalbfleisch, J. and Sprott, D.A., "Applications of Likelihood and Fiducial Probability to Sampling Finite Populations," in Johnson, N.L., and Smith, H., Jr., eds. New Developments in Survey Sampling, New York: John Wiley and Sons, Inc., 1969.

[3] Gonzalez, M.E., and Waksberg, J., "Estimation of the Error of Synthetic Estimates." Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, August 18-25, 1973.

[4] Royall, R.M. and Herson, J., "Robust Estimation in Finite Populations I," Journal of the American Statistical Association, 68, No. 344 (1973), 880-9.

[5] Royall, R.M., "Linear Regression Models in Finite Population Sampling Theory," in Godambe, V.P., and Sprott, D.A., eds., Foundations of Statistical Inference, Toronto: Holt, Rinehart and Winston of Canada, Ltd., 1971.

COMMENTS ON THE PAPERS BY GONZALES AND ERICKSON

Hyman B. Kaitz, C S R Associates

Both of these papers represent progress in the development of procedures for estimating local area data. Nevertheless, while they are work in progress, there is still some distance to go in attaining fully acceptable methodologies.

Both papers examine the accuracy of their techniques in terms of mean square errors of all the local area estimates, and in terms of the percentage of areas whose estimates deviate by more than a certain percentage, say 10%, from the known criterion values. It is desirable and relevant, of course, that such measures of dispersion or accuracy be used in judging these various techniques.

There is a particular problem, however, which arises in this connection. Estimates are produced for areas which are individually identified. For example, the method may yield a specific estimate for Altoona. Is Altoona interested in knowing whether its estimate comes from a body of area estimates with a satisfactory MSE, or is it interested in knowing how good the Altoona estimate is? I think the latter is more likely to be the case. In the absence of any alternative estimates, Altoona can perhaps become reconciled to its estimate, but it may not. It is interesting in this connection to see what has happened to local area population counts from the 1970 Census. Even though these are official government counts, they have been subjected to strong criticism by various local interest groups who have said that the presumed population undercount should be officially allocated to local areas differentially (black-interest groups), or that the Census Bureau should use more accurate enumeration techniques (Spanish-American interest groups).

When the data at issue are statistical estimates rather than official counts there may even be more opportunity for criticism. For example, local area unemployment rate estimates have been subjected to criticism for a number of years in states like California, Ohio, New Jersey and Massachusetts, principally because these estimates, based on administrative records from the unemployment insurance system, could be compared with the rates for the same areas based on the Current Population Survey. The prospect for critical examination of local area estimates is particularly strong when concrete incentives are present to seek estimates which increase local allocation of government funds under programs such as revenue sharing.

This suggests an approach to local area estimation which is not based on the use of a single technique, but on all the information available for a given locality which may help to improve its estimate. The amount of such information may differ from one locality to another. This would be a highly professional-labor-intensive activity and would probably be out of reach of most organizations with limited budgets.

With respect to the Erickson paper a question may be raised about the possible use of additional local area information available generally, which may be used as symptomatic variables, such as:

1) the racial mix in the base year
2) the urban-rural mix in the base year
3) the population density in the base year
4) The age-sex mix in the base year, and so on.

Erickson's work uses ratios as the symptomatic variables as well as the criterion variable. I would like to see some evidence on the amount of collinearity present among the symptomatic variables. In general, on statistical grounds, one seeks for symptomatic variables which are relatively uncorrelated with each other but are correlated with the criterion variable. These considerations should not, however, ignore those based on the subject matter under study.

I like Erickson's formulation of the regression equation to take account of the sampling error in the criterion variable. This assumes independence of the u and v error terms which does not appear to be quite true for his data set. This suggests extension of his model to include some nonzero but unspecified covariance between u and v. In addition, there may be reason to assume a v term for specific symptomatic variables, which would alter the model somewhat and would appear to produce biased estimates of the parameters of the regression equation in the ordinary least squares approach. There is also some heterogeneity in u, khich suggests seeking some transformation of the variables to correct this, or the use of generalized least squares in the estimation procedure.

Between the extreme of seeking a single estimation equation for all areas and the extreme of seeking to maximize the use of ad hoc local area information through a variety of techniques, there may be an optimum point at which clusters of areas may be studied, each cluster with its own estimation technique. Erickson's formulation suggests little payoff here if the analysis uses sample-based criterion variables.

Erickson's focus is on change over a long period of time. He suggests that for shorter periods, but still multiples of years, estimation may become somewhat more uncertain, and may call for the use of fewer symptomatic variables. Gonzales is concerned with the use of synthetic estimates for shorter periods of time, such as months or quarters. Here additional questions may arise with respect to volatile variables such as unemployment. For example, a local area may experience a sizeable layoff at a large plant, which would introduce an evident discontinuity in its time series, a discontinuity which would not

be reflected in the synthetic estimates tied to age-sex-color cells. I believe that the proposal to use occupation or industry cells which reflect changes in economic conditions more sensitively will do a better job, but that they will still reflect economic conditions in particular localities rather imperfectly.

When benchmark values are available, e.g. 1970 Census area estimates, it may be possible to apply a correction to the synthetic estimates. Let c_{io} be the Census estimate for the i-th area, and x_{io} be the synthetic estimate for the same area at that time. The difference, $d_i = c_{io} - x_{io}$, may then be taken as a first approximation adjustment for other time periods:

$$c^*_{it} = x_{it} + d_i$$

where c^*_{it} is the adjusted synthetic estimate. I expect that such an adjustment would help considerably in improving the synthetic unemployment rate for Honolulu, for which the latter is quite deficient. The use of other corrections should be explored as well, perhaps some which are functions of time or specific variables.

Finally, it is possible that synthetic and regression techniques may be combined in some way as to yield a combination superior to either alone.

EVALUATING THE EFFECTS OF INTERVENTION IN THE MILWAUKEE STUDY[1]

Howard Garber[2] and Rick Heber, University of Wisconsin

The research design of the Milwaukee study stands in contrast to previous longitudinal studies. Previous longitudinal studies have been either 1) descriptive, e.g., Fels and Berkeley Growth Studies; or 2) treatment oriented, e.g., Klaus and Grey, Weikart. The longitudinal studies in the first category were thorough descriptions, producing a large amount of correlational data, but were without a particular focus, perhaps because the population was not carefully selected according to set criteria. By considering a large number of variables simultaneously, this research was essential in establishing grounds and guidelines for later work.

The second group of studies was limited longitudinally, for their onset was not at the child's birth, i.e. they studied selected groups of children for several years or less. These studies have come under severe criticism because of their lack of adequate control. For the most part, the major selection criteria were low income of the families and age of the child, while maternal intelligence and a host of other important variables were not considered. Clearly, the focus of this second group of studies was remedial, not preventive. Often the treatment was short-term both in hours per day and in total duration. Specific program goals were sometimes lacking.

While the previous longitudinal research has provided segmental evidence of the importance of early development, it has never clearly coordinated the selection and the longitudinal aspects in such a way as to clearly evaluate development of a particular group of children as a function of a prescribed treatment.

The Milwaukee project was designed to determine whether "cultural-familial" or "socio-cultural" mental retardation could be prevented through a program of family intervention beginning in early infancy. This project differs from previous enrichment or intervention efforts in at least two ways. First, the subjects were selected on the basis of epidemiological studies which indicated that children born to parents who are poverty-stricken as well as of low intelligence are at high risk of being identified as mentally retarded. Secondly, the program begins in very early infancy and continues intensive intervention until the children enter first grade. The intention of this program is the prevention of mental retardation, in contrast to attempts aimed at remediation.

Before summarizing the results, I would like to review briefly the background and design of the Milwaukee project.

Approximately twelve years ago, the University of Wisconsin Research and Training Center established the High Risk Population Laboratory to study mental retardation among low income populations. The cultural-familial mentally retarded individual generally remains undetected until he enters school, since such mild intellectual deficiency is difficult to detect in the very young, especially when there is no evidence of organic damage.

The High Risk Population Laboratory is an area of Milwaukee, Wisconsin, previously found to have an extremely high prevalence of retardation, which we began to monitor by continuous door to door surveys. Though this area comprised about 2½% of the population of the city, it yielded approximately 1/3 of the total number of children identified in school as educable mentally retarded. According to U.S. Census Bureau data, the tracts comprising this area were in the lowest category of population density per living unit, percent housing rated as dilapidated, and unemployment.

All families residing in this area with a newborn infant, and at least one other child of the age of six, were interviewed and received individual intellectual appraisal. Through this survey we found clues for identifying families among the economically disadvantaged group with a high probability of producing a retarded child.

Specifically, we found a differential course of intellectual development for children born to mothers with different IQ levels. Furthermore, although there were no significant differences on the early infant intelligence tests between children born to mothers with above 80 IQ, and those born to mothers with below 80 IQ, after the infancy period, the children whose mothers had IQs greater than 80 maintain a steady intellectual level, while the children whose mothers had IQs less than 80 showed a marked progressive decline. (See Figure 1.) This trend toward a decline in measured intelligence for children in disadvantaged environments is widely accepted as a general characteristic of a "slum" environment population, yet these data indicate that the trend of declining intelligence with increasing age is restricted to offspring of low IQ mothers. In fact, we found the variable of maternal intelligence was the best single predictor of low intelligence in the offspring. The data indicated that the lower the maternal IQ, the greater the probability of the children scoring low on intelligence tests, particularly for the offspring of mothers with IQs below 80.

These observations from our survey data suggested our strategy to approach the prevention of socio-cultural mental retardation by attempting to rehabilitate the family rather than simply the individual retarded adult. The ability to select families "at risk" for mental retardation on the basis of maternal intelligence made it possible to initiate a program to study "high-risk" children before they become identified as mentally retarded.

As babies were born in our study area, trained surveyors employed by the University of Wisconsin Survey Research Center contacted the family within a few weeks of birth and completed a family history questionnaire which included a vocabulary screening test administered to the mother. Those mothers falling below a cut-off score on the vocabulary test were administered a full-scale WAIS by a trained psychometrist. A maternal IQ on the WAIS of less than 75 was the selection criterion in accumulating a sample of 40 families. These 40 families were assigned to either the Experimental or Control condition. It was not possible to accumulate at once a sample of 40 families where the mother met the WAIS selection criterion and then

randomly assign to Experimental and Control group because of the design requirement that intervention be initiated as early in infancy as possible. Our projections suggested that our screening procedures would identify about three families a month meeting criterion, requiring a little better than one year to accumulate our full sample. In actual fact, our projections were somewhat off; a total of eighteen months were actually required to generate the total sample.

Although this procedure constituted a deviation from strictly random assignment, it should be emphasized that only the happenstance of month of birth dictated group assignment. At no time did factors such as condition of the infant at birth, economic or domestic status of the family, etc., dictate group assignment. In fact, statistical analysis of differences in all measures present and known at time of birth, such as birth weight and height, recorded abnormality of delivery or condition of the infant at birth, marital status of family, economic status, and number of siblings were not significant. In addition, subsequent medical evaluations of the children as they grew have been carried out independently by staff of the Children's Hospital and Marquette Dental School. The analysis of these data revealed no statistically significant differences between groups in height, weight, serum lead levels, or other blood analyses.

Obviously, a major hazard for a longitudinal study of this kind is the potential for substantial attrition. We have been able to minimize this figure. Up to the present time, only two Control families have been lost and all efforts to locate them have failed. The Experimental group lost two subjects very early; one infant died as a result of a sudden crib death, and the second was lost by withdrawal of the mother from the program. This latter case represented the only instance of refusal to participate in the intervention program. Since both these losses occurred while the samples were still being accumulated, they were replaced, bringing the total N to 20; however, more recently, three Experimental families have been lost due to relocation to southern states. In two of these cases, the families left after the children had reached four years of age, and in the third case, after the child was 4½. Contact has been maintained with these three families however, and the children will receive the same comprehensive evaluations at the age of seven, as scheduled for all subjects.

The design of the Milwaukee project study for the Experimental group called for a comprehensive family intervention effort beginning in the home. The Experimental program was comprised of two components: (1) the infant, early childhood stimulation program and (2) a maternal rehabilitation program.

For the newborn infant, the program's objective was to provide intensive language and sensory-motor stimulation, and thereby facilitate the development of cognitive skills. Each day, beginning as soon after birth as was feasible - usually between three and six months of age - the child was picked up at home and brought to the Infant Education Center for the entire day.

The general educational program is best characterized as having a cognitive-language orientation implemented through a structured environment. Individualized prescriptive teaching techniques were utilized in the daily program (7 hours per day, 5 days per week). There was a high teacher-to-child ratio, which gave flexibility to the program and allowed for teacher feedback on the effectiveness of methods as well as individualization of instruction.

The program for the Experimental mother was designed to prepare her for employment and increase her awareness of her environment. This program included vocational training and classes in homemaking and fundamental academic skills.

The Control children, drawn from the same group of families as the Experimentals, were seen only for testing, which was done on a prescribed schedule for both the Experimental and Control groups of children. The testing schedule consisted of a comprehensive array of standardized and non-standardized measures of behavioral development, and was set from infancy to age seven where independent behavior evaluations are scheduled at the project's terminal point.

Our schedule of measurement included (1) developmental schedules of infant adaptive behavior; (2) experimental learning tasks; (3) measures of language development; (4) measures of social development; and (5) standardized tests of general intellectual functioning.

The Experimental and Control infants were on an identical measurement schedule, with assessment sessions every three weeks. The particular measures administered at a given session depended upon the predetermined schedule of measures for that age level. A particular test or task was administered to both Experimental and Control infants by the same person; the testers were not involved in any component of the infant stimulation or maternal program.

The Gesell Developmental Schedules were administered to both the Experimental and Control infants, beginning at age six months. Through the 14 month testing, the groups responded comparably on the four schedules: Motor, Adaptive, Language and Personal-Social. These data are represented as a composite of the four schedules, plotted with the mean scale developmental age norms for each age level tested. (See Figure 2.) At 18 months the Control group began to fall 3-4 months below the Experimental group, although still performing close to Gesell norms. At 22 months the Experimental group scores were from 4½ to 6 months in advance of the Control group on all four schedules while the Control group had fallen below the Gesell norms on the Adaptive and Language schedules.

Beginning at 24 months, increased emphasis was given to experimental, direct measures of learning and performance, as well as to the standardized tests of general intelligence.

The learning research program was designed to assess the longitudinal learning-performance characteristics of young children and to determine the role of these characteristics in the learning process. Furthermore, the role of this part of the assessment program was to provide more comprehensive information about cognitive growth than we were deriving from the IQ tests and various language measures.

We were concerned with delineating some of the characteristics of early learning behavior that either facilitate or interfere with performance. We wanted information on the response patterns or behavior styles, and how a child's simple response choice may reveal his general response tendencies and his ability to select and order incoming stimulation.

We employed a series of tasks including color form and probability matching. Our concern was with the child's strategy of responding: i.e., did he adopt a developmentally sophisticated strategy of consistent responding, either to color or form, or did he respond randomly, or perseverate to position? These learning measures have been administered every year since the children were 2½ years of age.

Our data revealed more developmentally sophisticated responding on all these measures by the Experimental group. Generally, the Experimental children have utilized a response strategy of altering successive responses according to the outcome of their previous responses. The Controls showed a tendency to perseverate on a response, e.g. to choose one position or to alternate from left to right, indicating that the children are insensitive to previous feedback and make no attempt to adopt a strategy.

We feel that in spite of the apparent simplicity of such tasks, they powerfully demonstrated the association of early intellectual development with the ability to impose order on the environment. Piaget (Inhelder and Piaget, 1958) suggests that response stereotyping is a manifestation of logical immaturity, and is a developmentally related deficiency in the use of higher order cognitive strategies. Even at five and six years of age, the percentage of Control children showing a tendency to perseverate was greater than the percentage of Experimental children showing such performance at three and four years of age. Thus, a response behavior which is important for future performance - the strategy or style of responding - appears to develop in the early years. The Control children strategies may interfere with their later learning while the style of the Experimental children should facilitate problem solving performance.

Our second major area of concern was the children's development of language and the measurement of this development.

The first statistically significant difference in language development appeared at 18 months on the Language scale of the Gesell Schedules. By 22 months, the Experimental children were over 4 months ahead of the norms and 6 months ahead of the Controls. This trend of differential language development has continued, in even a more dramatic way. In fact, some of the most striking differences in the performance of the Experimental and Control children are reflected in the research measures of language performance.

The analysis of free speech language samples indicated that the Experimental children between the ages of 1½ and 3 say more in conversation. Using this measurement technique, we find that it is not until three years of age, that the Control group produces a vocabulary comparable to that of the Experimental children. However, since the measure provides a gross picture of increasing language complexity, it actually masks the considerable linguistic differences that existed between the children. These differences show up in the group's performance on the more sophisticated language measures, beginning at age three.

At the age of three we began to test imitation with a sentence repetition test. This is an easily administered instrument which requires the child to repeat 34 sentences of varying length and grammatical complexity. The children's replies are analyzed for omission, substitutions and additions. The omissions are significantly greater for the Control group at every age level from 36 months on, while there is a significant decrease in omissions by the Experimental group every 6 months. Also, the Experimental group has substituted and made additions significantly fewer times in the repetitions. By the age of 4 the Experimental group made significantly more exact repetitions than the Control group, whose performance is comparable to the Experimental group's performance at 3. This same performance differential continues through age 5.

Also beginning at age 3 we tested grammatical comprehension with a modified version of a test developed by Bellugi-Klima (Fraser et al., 1963). This measure is a game in which the child manipulates objects in order to demonstrate his ability to understand 16 grammatical constructions. (The tester gives instructions for the child to fulfill a command, i.e. "Put the ball under the cup".) The results show that the Experimental group's performance is significantly superior at all age levels tested (3, 4 and 5). Their grammatical comprehension is one year, or more, in advance of the Control group.

Our standardized language instrument has been the ITPA, which has been administered to all children over 4½. The results have supported the differential performance of the Experimental and Control groups on our other measures. The mean psycholinguistic age of the Experimental group is 63 months (measured at 54 months) as compared to a mean of 45 months for the Control group: a difference in favor of the Experimental group by over a year and a half.

In describing the language behavior of the Experimental children, one would find them expressive, verbally fluent and according to the ITPA linguistically sophisticated. They speak their own dialect and they are proud of their own speech and yet their performance is developmentally advanced on sophisticated tests of the English language.

The next area we have given attention is mother-child interaction. We were concerned with the effects the intervention program may have had upon the family, particularly the mother. Previous research (e.g. Hess and Shipman, 1968) found that the mother's linguistic and regulatory behavior induces and shapes the information processing strategies and style in her child and can act to either facilitate or limit intellectual growth.

In the mother-child interaction most sophisticated behavior - such as the initiation of problem-solving behavior by verbal clues and verbal prods, or the organization of tasks with respect to goals in problem-solving situations,

etc. - is done by the mother. However, where the mother has low IQ, the interaction is more physical, less organized and less direction is given to the child. Indeed, while this was the case in the Control group mother-child dyads, it was quite different in the Experimental dyads.

We found that the Experimental dyads transmitted more information than the Control dyads, and this was a function of the quality of the Experimental child's verbal behavior. The Experimental children supplied more information verbally and initiated more verbal communication than the Control dyads. The children in the Experimental dyad took responsibility for guiding the flow of information - providing most of the verbal information and direction. The mothers of both dyads showed little differences in their teaching ability during the testing session. However, in the Experimental dyads, the children structured the interaction session either by their questioning or by teaching the mother. Also, the Experimental mothers appeared to be modelling some of the behaviors of their children. Consequently, they used more verbal positive reinforcement and more verbal responses.

As a result, a developmentally more sophisticated interaction pattern has developed between the Experimental children and their mothers, which contributed to faster and more successful problem completion.

It became apparent from these data of the mother-child interaction, that the intervention effort has effectively changed the expected pattern of development for the Experimental dyads. Moreover, the result of what might be termed a reciprocal feedback system initiated by the child has been to create a more sophisticated, more satisfying interaction pattern in the Experimental dyad. In fact, there is some evidence that the Experimental mothers might be undergoing some changes in attitude and self-confidence. The Experimental mothers appear to be adopting more of an "internal locus of control" - an attitude that 'things happen' because of their decisions and actions and not purely by chance or fate. Thus, the intensive stimulation program, in which the Experimental children participated, has benefited both the Experimental child and the Experimental mother by broadening their verbal and expressive repertoire.

A clearer picture of the differences between groups is given by the results from standardized measures of intelligence.

We have presented the summary data from intelligence testing in Figure 3.

We have derived data from 12 to 21 months from the Gesell Developmental Schedules. The standardized intelligence scores at 24 months are from Cattells and from Binets, thereafter.

As you can see, the mean IQ of the Experimental group is consistently 25 to 30 points above that of the Controls. For example, at 60 months the mean IQ of the Experimental group was 118 in comparison to the Control mean IQ of 92, a difference of 26 points. We have calculated IQ at the 72 and 84 month points, but they include scores from less than the complete group of Experimental or Control children. These points are particularly important for they are some of the first evaluative data obtained since the children have been out of the intense education program and on their own. Although there was a drop in the scores of the Experimental children to 112 at 72 months, and 110 at 84 months, where some of the children have completed first grade, there is a comparable drop for the Controls, lowering the mean score to 87 at 72 months and 84 at 84 months. It is particularly significant that with the decline in test scores, there has remained a large differential in mean IQ, which the Experimentals have maintained throughout the testing: at 84 months there is still a 30 point difference between the groups. We are encouraged by these preliminary results for if you remember, the purpose of this program was to prevent a decline in intellectual functioning with age increases to the retarded mean IQ level of their siblings and mothers. This decline is in evidence for the Control group, whereas it seems likely that the Experimental group will level off at mean IQ level about 100. Of course, next year's testing will give us a more complete picture of the progress of both groups.

The tendency for declining IQ in this population is further illustrated by the comparison data in Figure 4. The bottom dotted curve is the original survey group. The longer solid line curve is the mean IQ of the siblings from both the Experimental and Control families. In general these older siblings of our actual subjects show the same trend toward declining IQs with increasing age, as do the actual Control children, whose mean IQ data is represented in the shorter solid line curve. Thus, it appears that we have prevented in the Experimental group the relative decline in intellectual development that we see now in the Control group, and that we found in the siblings of both groups and in the original survey groups.

I think these data answer one of two pivotal concerns about the study at this time. One concern, obviously, is the basis for predicting that these children are at risk for retardation. In other words, can we be sure that the downward trend in IQ for this population is reliable? The data I have just shown indicates that it is; successive samplings from four generations of offspring have shown the same tendency to declining IQs. The second major concern at this time is whether the present differential performance favoring the young Experimentals is merely an artifact of training. The strength of the present differential performance in favor of the Experimental group is borne not only by their standardized test scores and their performance levels on the various experimental tasks, but also by the differential behavior patterns displayed by the two groups. The pattern of the Experimental children indicates a sincere and concerned effort to work the task, while the Control children have tended to be apathetic and perseverate their response.

Even with such a comprehensive assessment program, interpretations of, and generalizations based upon present data must be tempered not only by recognition of the test sophistication which has obviously been acquired but also by knowledge of previous enrichment studies where treatment gains have not been maintained over long post-treatment periods. We have planned independent,

comprehensive behavioral evaluations to be conducted a year beyond the termination of intervention. These data may prove a more reasonable basis for evaluation of effects on intervention. This is not to suggest that subsequent changes on relative performance levels would not occur beyond that level, but rather, it would provide a more solid basis for evaluation of the treatment effects. Any ultimate evaluation, of course, must be based on the performance of these children as they move through the educational system. We are encouraged by the preliminary results from the members of the Experimental group who have completed first grade. However, final interpretations will be put off until more of the group have reached this point.

We shall continue to be quite cautious in the interpretation of our data. This is not peculiar, particularly when one considers the numerous pitfalls and hazards of infant measurement. The Experimental children have had training, albeit fortuitously, on items included in the curriculum which are sampled by the tests, while the repeated measurements have made both groups test-wise. We have tried very hard to answer whether it has been simply a matter of training and practicing specific skills. In fact, extraordinary precaution has been taken to separate the development of the curriculum and the assessment program. Two separate staffs have been employed. It is obvious to most researchers that, to some extent, infant intelligence tests must contain material which approximates material used in preschool curriculum, primarily because of the limited variety of material for this age. To circumvent this problem somewhat, we employed other measures of performance, which minimized the stock item, and thereby afforded additional insight into the differential development of these children. As could be seen in the measures of learning and language development, the differential performance discrepancy is consistent with the IQ measures, indicating advanced intellectual development of the Experimental group. What's more, there is considerable difference in the pattern or style of behavior between the groups -- particularly the tendency to stereotype in the responses exhibited by the Control group, which certainly is antagonistic to successful learning performance.

[1]Research supported in part by Grant 16-P-56811/5-09, formerly RT-11, from the Rehabilitation Services Administration of the Social and Rehabilitation Service of the Department of Health, Education and Welfare.

[2]A more comprehensive report is available from the first author at the Waisman Center on Mental Retardation and Human Development, 2605 Marsh Lane, University of Wisconsin, Madison, Wisc, 53706.

Fig. 1 IQ Decrements in Disadvantaged Children Whose Mothers are Mentally Retarded

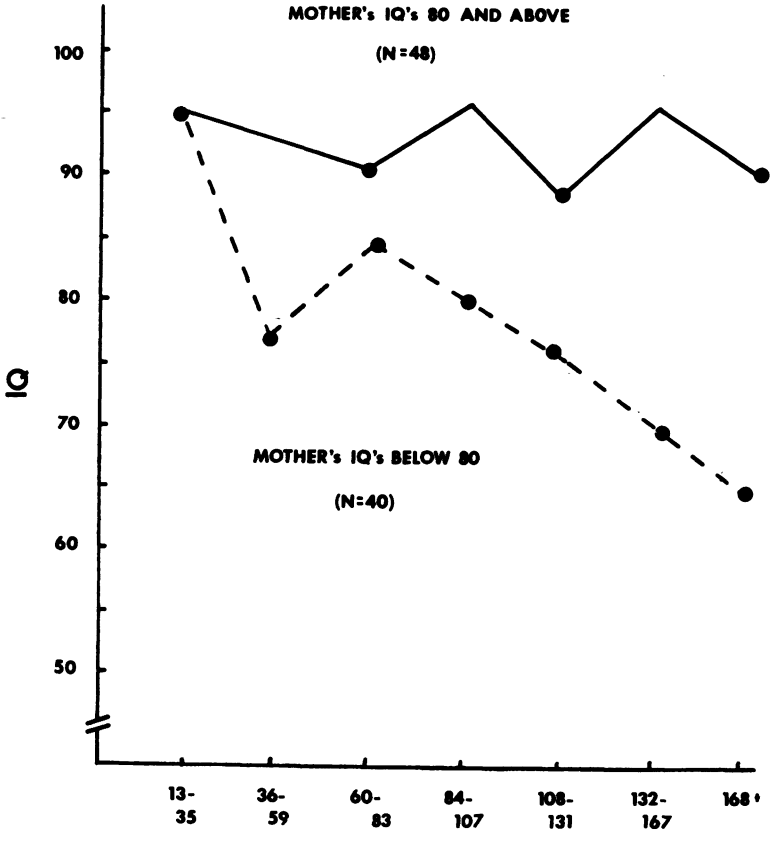


Fig. 2 Composite of 4 Gesell Scores

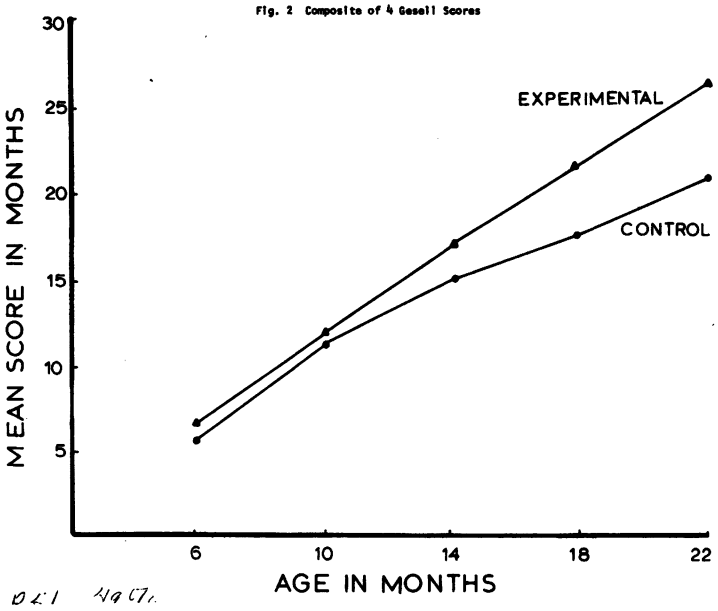


Fig. 3 Mean IQ Performance with Increasing Age for the Experimental and Control Groups

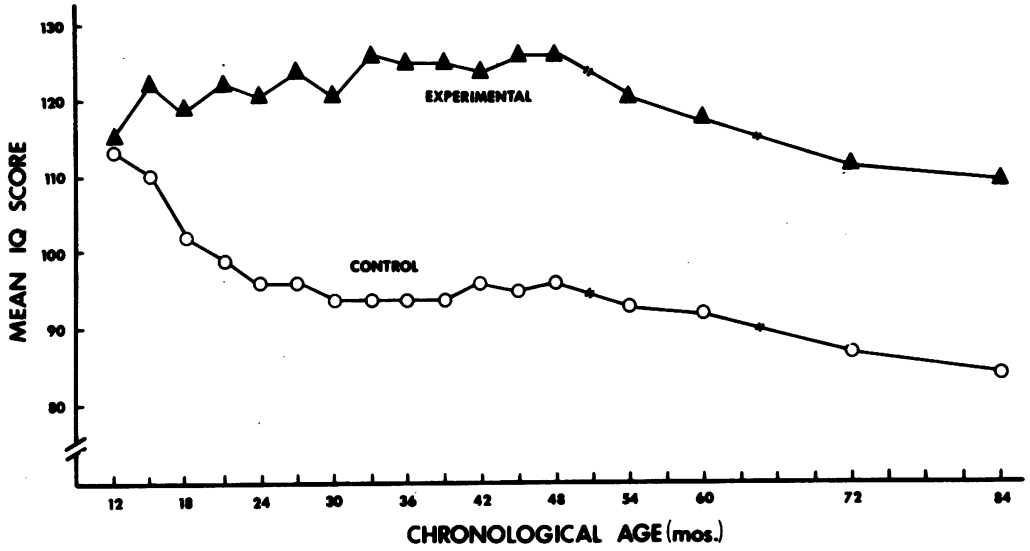
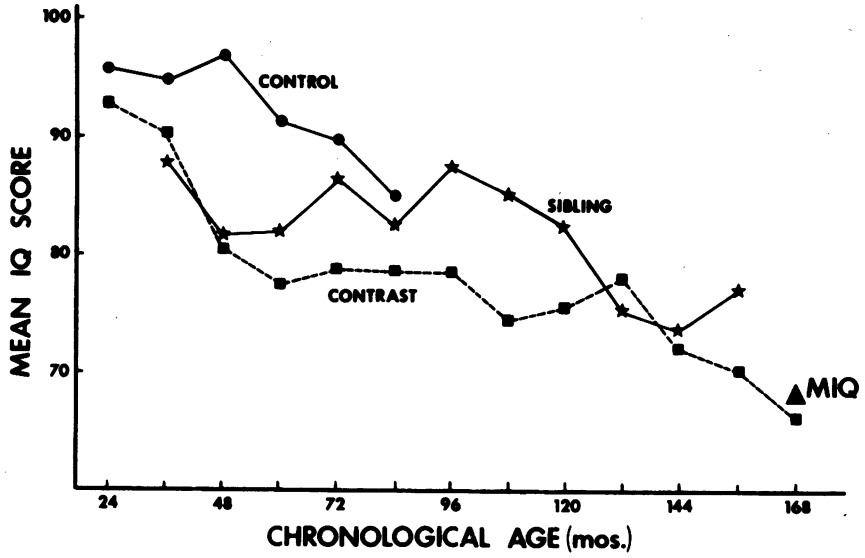


Fig. 4 Sibling and Contrast Group IQ Scores



DISCUSSION

Craig T. Ramey, Frank Porter Graham Child Development Center, Univ. of North Carolina

As Cronbach (1957) has pointed out some years ago, research within the domain of the social and behavioral sciences tends to be either correlational in nature or experimental in approach. Long and productive debates have raged concerning which of these approaches generates the more useful kinds of data with which to begin to alleviate major social problems. At the one extreme, complex social systems defy adequate description when elementary, univariate statistics are used as decision-making criteria. Therefore, in many cases, it appears that the best that the bio-social scientist can do, when dealing with large social issues is to find patterns that can be extracted from matrices of intercorrelations of many variables. Yet as every elementary student of statistics knows, the dictum that correlation does not imply causation serves as a continual bane to the full understanding of the directionality of social forces. As a consequence, there are strong arguments for developing an experimental approach to the understanding of human behavior even within the context of complex social systems. Yet as Heisenberg (1962) has pointed out so clearly, the very process of measurement distorts the phenomena to be observed.

A frequent suggestion to those concerned particularly with the understanding of the development of human behavior within complex social systems is to study the development of groups of individuals over time; i.e. to begin a longitudinal investigation of development which will allow one to describe not only the behavior of groups of subjects and the impact that various environmental influences may have on that group's development, but also a longitudinal investigation allows one to examine the course of development of specific individuals within that group. Parenthetically, educators are now strongly suggesting that we spend more time understanding our treatment X subjects interactions than we have historically.

While, in principal such an approach seems to be a reasonable and even highly a desirable method of procedure, the implementation of adequate research designs to understand highly complex social phenomena is extremely difficult-at best, and at worst impossible--given our current state of knowledge for decision making rules and techniques for drawing causal inferences. With respect to the desirability of longitudinal research in general, as compared to cross-sectional approaches, it must be pointed out that several large and torturous traps lie in the path of even the most dedicated and competent of researchers. Again, it is elementary to point out that one's findings truly do not generalize beyond the population from which the sample has been selected. Thus, a study performed in Milwaukee or in Chapel Hill, or in New York City, or in Berkeley does not of necessity generate data on human development which is directly extendable to other communities or to other countries; yet, one hopes that certain basic principles of development are so general and so pervasive in nature that one may safely and practically proceed to make such generalizations even though he realizes that in theory he must do so with extreme caution.

A morass of other problems plague the researcher who begins a longitudinal study and many of these have been pointed out in an excellent and thoughtful series of methodoligical papers by Warner Schaie (1965) and his collegues including Paul Baltes and John Nesselrode, among others. However, much as these problems are with us, we are left on the horns of the dilemma. Do we seek not to engage in broad scale research which can in fact generate information relative to the human condition, because we dare not violate serious assumptions? Or do we recognize that in many cases we are violating the assumptions and proceed to make the best analysis that we can of our rather imperfect data?

Rick Heber and Howard Garber and their staff are to be highly commended for having the audacity to design an experimental program which seeks to resolve a large part of the ubiquitous nature-nurture controversy. They are particularly to be applauded for attempting to resolve this problem, not within the hothouse of the sheltered laboratory (which would be hard and meaningful enough if successfully accomplished, even if the generality of the finding were limited) but, rather, they have sought to generate data relevant to the heredity-environment controversy within the most meaningful of all possible realms - the realm of real life, with real people, with very real problems. Further, those problems are embedded within a social context which itself is nearly impossible to comprehend and to describe precisely. Thus, in all sincerity we must praise the magnitude of the effort and, because the problem is so vitally important, we must also bring our best acumen to bear upon the adequacy of their methodology, the precision of their data analyses, and the validity of their conclusions.

Methodology

The primary justification for their subject selection rests upon the finding from a cross sectional design study which purportedly shows that the IQ's of children whose mothers score below 80 on an IQ test show a progressive decline in intellectual performance when the children are between 13 and 168 months of age. Warner Schaie (1965) has sufficiently cautioned us against such straightforward extrapolation of longitudinal trends from cross sectional analyses that we must all realize the fragileness of such extrapolation. Although the papers from the Heber-Garber project are not written in sufficient detail concerning the sampling and testing procedures

for me to fully evaluate the validity of their conclusion, it seems to be a plausible one. In any event, for the sake of discussion I will tentatively accept their findings as valid.

As a result of this finding the authors selected families for inclusion into their intervention program whose mothers scored less than 75 on an intelligence test. Forty such families were identified and assigned either to an experimental or a control group. The number of families contacted, the refusal rate, the demographic differences between those selected and those who accepted as far as I can determine have not been presented. The method of assignment of these families is equally crucial for the interpretation of their findings. Yet from the papers that I have had access to, there simply is not enough information to evaluate the adequacy of their assignment procedures. Were the families randomly assigned? Were they matched on a priori determined variables and then randomly assigned? Or was some other procedure employed? It is imperative that we know the answers to these questions because almost all subsequent statistical analyses make assumptions in this area. The consequences of violating these assumptions in any particular sample simply cannot be determined.

Data Analyses

Although one wants to ask many detailed questions about specific data analyses, there appears to be one pervasive issue which leaves one unsettled. The essential nature of this project is to take many measures on the same individuals over time in what is essentially a repeated measures design. Although it is not spelled out in the written version of Dr. Garber's paper, I presume that Winer (1971) type analyses of variance with repeated measures formed the bulk of their analyses. The criticism which I am about to make must be understood in context. It has been seven years since this project was begun and many statistical innovations and refinements, thanks to many of you in this audience, have been made in the meantime. Nevertheless, if such analyses were performed, they must be reevaluated in light of recent criticisms by McCall and Applebaum (1973) among others. For example, unless one can assume homogeniety of the covariance matrices in a repeated measures design, then it is likely that one inflates his alpha level beyond what he has actually chosen. Various methods of protecting one from resulting type I errors exist to minimize such risks. For example, one can use the Box (1954) correction for the degrees of freedom in univariate analyses or one can use the MANOVA program with the Wilks Lamda Criterion which does not require an assumption of homogeneity of covariance. Several other analysis questions could be offered but I shall limit myself to these two because of time constraints.

The Intervention Program Itself

Seven years ago the prevailing professional opinions on the expectations of the possible outcomes of intervention programs were divided into basically two different camps. On the one hand there were those who felt that specific deficits in language and cognitive development could be remediated through some relatively short term compensatory experiences which would significantly enough alter the course of intellectual development, such that children who were already developmentally retarded would be remediated and innoculated against failure in future academic settings.

On the other hand there were those who felt that only by a drastic alteration of the base and deplorable social forces which press unbearably upon the oppressed and the poor, could the bright futures, which socially conscious Americans advocate as the birthright of all citizens, be achieved.

The disappointing results from the short term remedial intervention attempts lay waste to the ideal of a simple solution to a complex social problem. Even before the Headstart results were tabulated, the Milwaukee group was quietly moving toward the enormous and consuming task of beginning at the beginning. An in medias res solution was given way to a reexamination of genesis. With little in the way of articulated theory of early and global human development and with even less help in the form of curricular products from educators, the bold conception of starting essentially at birth was born. As one who is now currently trying to evolve a theoretically generated curriculum for ensuring optimal early development, I can more than sympathize with Heber's often quoted and much criticized statement that the Milwaukee curriculum consisted of every thing but the kitchen sink.

Because the curriculum and the other components of the project are essentially the independent variables in this and other intervention programs, we, as scientists, are left with feelings of dissatisfaction about the project and its replicability. Harsh and severe criticism can and will be levied against the imprecision of the specification of the treatment variables. It is, and will continue to be, nearly impossible to determine which variance components have been accounted for by the improved nutrition which the experimental group undoubtedly received; by the potential Hawthorne effects associated with being in the experimental group and by the myriad other variables which might have influenced the differential performance of the two groups.

Such criticisms can and should be made. They should be made because only through such a process of critical examination will the truly important manipulable variables be isolated and examined. Yet one should temper one's criticisms with the realization that science proceeds from the molar to the molecular levels of explanation. Newton's work had to preceed Einstein's and Einstein's had to preceed Heisenberg's. So too we in the bio-social sciences can and must proceed.

We now know empirically through projects such as this one that the course of human behavior can now be profoundly altered; that developmental

retardation is not the birthright of even our most oppressed and disadvantaged children. What remains to be done now is to move from the level of description of the consequences of social and individual change to an understanding of the processes and mechanisms whereby such change is theoretically understandable and exportable for the elimination of preventable retardation.

References

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, II: Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics, 1954, 25, 484-498.

Cronbach, L. J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.

Heisenberg, W. Physics and philosophy: the revolution in modern science. New York, 1962.

McCall, R. B. and Appelbaum, M. Bias in the analysis of repeated measures designs: some alternative approaches. Child Development, 1973, 44, 401-415.

Schaie, K. W. A general model for the study of developmental problems. Psychological Bulletin, 1965, 64, 9 2-107.

Winer, B. J. Statistical principles in experimental design. New York: McGraw Hill, 1971.

DISCUSSION

Richard J. Light, Harvard University

The paper delivered by Professors Heber and Garber raises a number of fascinating questions about how to interpret the evaluation of a social intervention. I would like to devote my discussion to several of these points.

First, the investigators should be strongly praised for conducting this intervention as a true experiment, using randomization of subjects between treatment and control groups. Nearly all work in preschool intervention has not been experimental in the statistical sense; rather, programs have been evaluated by post hoc studies. Such studies usually have to create an after the fact artificial control group consisting of children with similar characteristics to those in the experimental group. Thus, because randomization was not employed, one rarely has any serious degree of confidence in the results of these studies. The Heber and Garber work obviously involved a serious effort to carefully define the study population of interest, and then to randomize children into two groups.

My major criticism of this work centers around how the treatment was defined. On page 6 of their paper, the authors report that, "individualized prescriptive teaching techniques were employed." On one hand, that is a constructive approach, since no doubt the families in the study displayed a range of individual differences in the kinds of help they would benefit from most. But on the other hand, the lack of a precisely defined treatment limits very severely the generalizability of any positive results. To be specific, suppose the city of Cambridge, Massachusetts decided that since the results of the Milwaukee investigation were so promising they wished to institute the treatment. What should they institute? I am afraid that the answer is not forthcoming from the research report of the Milwaukee Project. To say that help should be offered to mothers with low IQs on an individual basis, depending upon their needs, does not clearly specify any treatment. Thus the external validity, or generalizability, of results is severely limited.

A similar difficulty is created for the investigator's ability to make inferences internal to the study. The results reported by Heber and Garber can be summarized as being essentially a smashing success. Intervening with low IQ mothers and their children seemed to lead to enormous IQ gains on the part of the children, relative to a randomized control group. But what precisely was done with these mothers and children? According to the presentation, different families had different forms of intervention, with stress on different kinds of facilities. Thus, if we had to attribute the enormous program success to a particular treatment feature, we might easily conclude the critical feature to be the sensitivity and excellence of the teachers and social service people employed in this project. It is a pleasure to congratulate these people on their excellent work. But, once again, the difficulty for other investigators is that a treatment defined as being "unique for every family" is not a treatment easily generalizable, except perhaps in the extremely limited case where the identical social service team is involved.

This question of treatment specifiability is my primary criticism, and should be viewed in the light of my earlier comments about how delighted I am to see a randomized study of this nature. Let me now move ahead to a series of short questions about other issues raised by a preschool intervention study of this kind.

The primary dependent variable in this study was IQ. The treatment group had IQs after several years of intervention that far exceeded those of the control group. But the value of IQ scores is, ultimately, their ability to predict reasonably well a child's later performance in school. It will be interesting to see whether the higher IQ scores achieved by the treatment group children are good predictors of their school performance. In other words, there is some chance that these elevated scores, while correct in that they were obtained honestly, have a lower degree of predictivity than do IQ scores of children who have not been in a treatment group. A very likely possibility is that the treatment children will do better in school than the control children, but nowhere near as much as their remarkably increased IQ scores suggest they might.

The paper presents mean IQ scores for the two groups over a period of five years. But no data on variance is given. I am sure that Professors Heber and Garber have these data, and I raise this question not as a criticism, but rather out of curiosity as to whether some subject by treatment interactions may exist in these data. That is, might it be possible that "the program" is superb for some children of low IQ mothers, but not for others? Is it further possible that we might be able to identify what kinds of children, with what kinds of features, benefit most from this Heber and Garber program, so that it might be optimally targeted to children who are most likely to benefit? Such a result would not be surprising or unique, as much research with Head Start, the preschool intervention program, has shown that different kinds of curricula differentially benefit different kinds of children.

I would like at this point to mention one substantive finding of a colleague at Harvard,

Professor Burton White. In working with middle income children, whose mothers have IQs in the "normal" range, he has reported that the critical period for children developing skills that tie in with later school competence is the age range 10 to 18 months. Professors Heber and Garber's data indicate that the critical period when the control group children begin to separate from the treatment group children first occurs at about age 18 months. I wonder if this difference is due to some feature of the treatment being reported here, or rather might have something to do with the social class of the families being studied. I have no idea as to the answer but would be interested if the investigators felt this difference was due to social class differences in the families.

A last observation has to do with the feasibility of adopting the form of family intervention reported by Professors Heber and Garber on a widespread national scale. The exciting feature of their work is that if their results hold up in replicated studies elsewhere, they will have succeeded in demonstrating that it is possible to substantially increase the IQs of children by an intensive intervention program begun at a young age. The worrisome feature is the practical one of cost. While no precise dollar figures are reported in their study, it is clear that the cost of the intervention runs to several thousand dollars per child per year. It is necessary then to ask the question, how does this intervention stack up on a cost benefit comparison with other forms of preschool remediation, such as, for example, the television program Sesame Street? I believe that the cost of Sesame Street is less than one dollar per year per child. Without arguing that televised instruction confers greater relative benefits than the intensive intervention program discussed here, I believe it is necessary for policy makers to raise this question.

To tie up these comments, I believe that the report by Professors Heber and Garber has two clear strengths. First, it illustrates that intensive remediation offers hope for substantially improving the IQs of children from low IQ family backgrounds, and this is most promising. Second, by using a randomized approach, it gives us greater confidence that the positive findings are not due to some artifact such as self selection, which would ultimately negate the value of these optimistic findings.

On the less optimistic side, the lack of a clear specification of the treatment, and the fact that the treatment varied quite a bit from family to family, makes it very difficult to generalize the results of this study. Science depends upon a steadily accumulating body of evidence from which, over time, we can draw stronger and stronger inferences. But it is very questionable whether a treatment that is not precisely defined can be replicated in different times and in different places, and this difficulty holds up scientific progress. Finally, I would urge that Professors Heber and Garber, together with other investigators who may decide to extend these findings, examine particularly carefully the question of subject by treatment interaction. A growing body of evidence in preschool education suggests that different types of curricula are best for different kinds of children. Discovering these interactions enables us to direct particular programs to the children who are most likely to benefit, and thus helps us organize effective social policy.

THE MEASUREMENT OF CRIME THROUGH VICTIMIZATION SURVEYS: THE CENSUS BUREAU EXPERIENCE

Marie G. Argana, Marvin M. Thompson, Earle J. Gerson, U.S. Bureau of the Census

Definition and control of criminal behavior historically have been functions of the State. Accordingly, governments have long compiled information on various aspects of crime, including law enforcement, court actions, and corrections. Such information is used to determine the extent of crime and the effectiveness of control procedures.

In the United States, information on trends in the amount of crime is derived from administrative records of law enforcement agencies, which are largely units of local governments. On the basis of reports from these police departments, the Federal Bureau of Investigation compiles national summaries entitled the Uniform Crime Reports (UCR).

The UCR utilize seven crime classifications to establish an index to measure the trend and distribution of crime. The crimes selected--murder, forcible rape, robbery, aggravated assault, burglary, larceny over $50 and auto theft--represent the most serious crime problem. However, one major limitation of these data is that many crimes are not reported to the police. Another difficulty with the current data is that administrative records necessarily provide a limited amount of information about the event, the victim, and the offender. These records do not uniformly provide the kinds of information necessary to an understanding of crime extending beyond its incidence and type to such aspects as the characteristics of the victim and offender and a detailed description of the event.

In 1965, the U.S. President's Commission on Law Enforcement and Administration of Justice was established to inquire into the causes of crime and delinquency and to make recommendations for its prevention, as well as the improvement of law enforcement and the administration of criminal justice. As part of the Commission's work, the first nationwide survey of crime victimization was initiated. The National Opinion Research Center of the University of Chicago surveyed 10,000 households to determine if any household member had been victimized, if the crime had been reported to the police and, if not, why not. This study concentrated on the same crimes as reported in the UCR. More detailed surveys about the UCR crimes were undertaken in a number of precincts in Washington, Chicago, and Boston by the Bureau of Social Science Research and the Survey Research Center of the University of Michigan. These studies all indicated that the amount of actual crime was much greater than that reported in the UCR and further, that a great deal of crime is unreported to the police.

In 1970, the Law Enforcement Assistance Administration (LEAA) asked the Bureau of the Census to begin developmental work toward a national sample to provide victimization data through household surveys. The first studies employed a reverse record check technique in which victim respondents were identified from police records. The primary objectives of these studies were to determine the reference period about which to question the respondent to gain the most complete and reliable information; to measure the degree of telescoping; i.e., the tendency of the respondent to advance an incident occurring outside the reference period into that period when questioned; to explore the possibility of identifying incidents by a few broad general questions as opposed to a series of more specific probing questions; and to test and improve the survey instrument. The respondent was asked a series of questions on a screening questionnaire to determine whether he had been victimized. This was followed by a detailed incident report designed to classify the crime and obtain additional information about the incident.

A certain amount of telescoping (placing the incident within the time frame, when actually it occurred before) did appear. Generally, when this occurred, the incidents were reported to have taken place 1 or 2 months prior to the reference period. This telescoping tendency can be controlled by using a bounded interview technique. With bounding, information obtained in a previous interview is used to prevent duplicate reports.

The first nationwide household survey effort was conducted as a supplement to one of the Census Bureau's household surveys, the Quarterly Household Survey (QHS). At the time this work was undertaken, the sample was spread over 235 primary sampling units (PSU's) throughout the United States, and housing units in the sample were interviewed for six quarters, with one-sixth of the sample retired and a new sixth introduced each quarter. Interviews on crime victimization were conducted in approximately 15,000 occupied housing units. These supplements, conducted at 6-month intervals from January 1971 through July 1972, were used to test and improve the questionnaires and accompanying field procedures and to revise and improve the clerical and computer tabulation procedures. They were also used to address certain methodological issues such as the use of a mail screening procedure, the control of telescoping through the bounding technique, the use of the telephone to obtain details of reported incidents and the advisability of lowering the minimum age of coverage from 16 to 12.

In addition, the Bureau also conducted victimization studies in two cities (San Jose, California, and Dayton, Ohio) that had participated in LEAA's Pilot Cities Program, which provides funding assistance to law enforcement agencies. These studies were also used for methodological research. One of the most important methodological issues addressed in the Pilot Cities Studies was the use of self-respondent versus a household respondent. In all

of our previous work (including the QHS supplements) one household respondent answered all the screening questions and detailed incident reports for all eligible household members. In the Pilot Cities Surveys, the household respondent technique was used in one-half of the households in each city, while the self-respondent approach was used in the other half; i.e., all eligible household members responded to the screen questions and the incident reports for themselves. A series of attitudinal questions concerning such topics as attitudes toward neighborhood, local police, and fear of crime also was included.

This developmental work led to a series of conclusions which directly affected the major effort that followed. These conclusions are as follows:

- the optimum recall period was either 3 or 6 months.
- telescoping appears to be effectively controlled by the bounding technique.
- the self-respondent technique elicited more reports of victimizations than did the household respondent technique.
- a series of detailed screening questions were more effective than several general screening questions.
- a single incident report, covering both personal and property crimes was developed.

Using much of the information obtained in the aforementioned studies (which were related to household surveys only), a continuing survey, the National Crime Panel, was established in July 1972 covering a general probability sample of households and commercial establishments. The commercial victimization survey, which parallels the household survey in many respects, is not described in this paper.

The National Crime Panel has two major components, a national sample and a sample of large cities. The national household sample consists of approximately 72,000 designated housing units in 376 primary sampling units. The sample is divided into six parts, each of which is interviewed in a specified month and again at 6-month intervals; i.e., July and January, August and February, etc. Households included in the national sample are interviewed seven times. Approximately 12,000 new households are introduced over each 6-month period, replacing approximately the same number which rotate out of or leave the sample during the 6-month period. There are several reasons for rotating the sample, one of which is to avoid the loss of cooperation which may result from interviewing the same households indefinitely. Another is to reduce the effect of possible biases in responses when the same persons are interviewed for an indefinite period.

The second component, the cities sample, consists of approximately 12,000 housing units in each of 30-35 central cities in the largest Standard Metropolitan Statistical Areas (SMSA's). These include the five largest cities in the United States (Chicago, Detroit, Los Angeles, New York, and Philadelphia), as well as others selected for varying reasons (participation in special LEAA programs, cities with particular crime problems, etc.)

Interviewing in the five largest cities was conducted at the beginning of 1973 and will be repeated early in 1975. Other cities will be interviewed at 3-year intervals. Because households in the cities sample will be interviewed only two to five times during the decade, no rotation of sample households is planned.

Data collected in the initial interview for a household in the national sample are not used in preparing final estimates, but rather serve a bounding function; that is, to ensure that earlier incidents are not telescoped into the successive visit. The effect of telescoping on the level of reported crime will be studied during the first few years of the survey by comparing first-time unbounded interviews with data from subsequent bounded interviews.

The sample will not be completely bounded until incidents that occurred in July through September 1973 are recorded. Preliminary information indicates that this bounding technique will affect these data. Personal crimes were reported at a .730 ratio of bounded to unbounded households, while the ratio for property crimes was .765.

The bounding technique is not used in the cities sample because of the gaps in the time coverage of the survey. In the five largest cities, for example, the sample was first interviewed in 1973, essentially covering victimizations in 1972. Interviewing in 1975 will cover 1974 victimizations. Without data for 1973, it is impossible to bound the 1974 data. However, we do plan to develop bounded estimates for the largest cities from the national sample and compare them with the unbounded city results to provide some measure of the possible effect of bounding.

As was noted earlier, one of the important problems in the National Crime Panel is victim recall. Results of the earlier studies indicate that the shorter the recall period, the better the respondent is able to remember any victimizations that occurred during that period and to remember when they occurred. For the national sample, the reference period was set at 6 months, ending the last day of the month prior to the interview month; i.e., households interviewed in July 1972 (the initial interview) were asked about victimizations that occurred during the period from January 1, 1972 to June 30, 1972. The reference period used in the cities sample is a variable 12-month period, ending the last day of the month prior to the

interview month. One reason for doing this was to provide both the respondent and interviewer with some fixed point in time so that all data collected during a particular month would refer to the same period.

In the national sample, incidence of crime can be measured over a period of time, but in the cities sample, this incidence can be measured only for a specific period of time or at a specific point in time. The cities data may be used to establish baseline estimates to be compared to estimates produced at a later point in time. Many of the cities will be using these data in exactly this way. The baseline data will provide an estimate prior to the implementation of programs designed to reduce certain aspects of crime and the later estimate can be used to measure the effect of these programs.

Another issue raised during the developmental phase was whom to interview; that is, should one household member provide information regarding all the victimizations of all eligible household members or should each eligible person be interviewed for himself, regarding only his own victimizations? The Pilot Cities studies indicated that there was a significant difference in the victimizations reported by the household respondent compared to the self-respondent. As might be expected, many more victimizations were reported by persons reporting for themselves than by the household respondent method. On the basis of this evidence, the self-respondent technique for all persons 14 years and older was adopted. Information for persons 12 and 13 years old is obtained by interviewing another adult household member (usually a parent).

The household portion of the National Crime Panel focuses on measuring the extent of victimization ascribable to the major index crimes of assault, burglary, larceny, auto theft, and robbery. A series of screen questions are asked to determine if any attempted or actual victimization occurred during the reference period. After all the screen questions have been completed for an individual, questions designed to obtain information on the circumstances and characteristics of the incident are completed for each reported incident. These include items such as time and place of occurrence, injuries suffered, medical expenses incurred, number, age, race, and sex of offenders, relationship of offender to victim (stranger, casual acquaintance, known by sight, relative), and other detailed data relevant to a complete description of the incident. In addition, data are being collected about the victim on such subjects as education, migration, labor force status, and if employed, occupation and industry in which employed.

Legal and technical terms, such as assault and larceny are not used in the interview. Rather, through a structured questionnaire, a complete description of the incident and the elements of the behavior (both victim's and offender's) is obtained. On the basis of the description, the incidents are classified at a later time into the index crime categories or are excluded from the survey as out-of-scope.

Interviews in the national survey are conducted by the Bureau's staff of permanent interviewers. Interviewers are given an extensive initial training program, which covers material relevant to the general current surveys conducted by the Bureau, as well as that which is concerned solely with the National Crime Panel. Monthly memoranda, explaining new procedures or clarifying existing procedures, are sent to the interviewers. Periodically, the interviewers are required to attend refresher group training programs.

An interviewer observation and reinterview program, to ensure the quality of data collection, is an integral part of the field operation. Supervisory personnel observe each interviewer on a regular and continuing basis throughout the year. In addition, each month a sample of the interviewers' work is reinterviewed by supervisory personnel. Based upon the observation and reinterview results, interviewers identified as needing help may be retrained in particular aspects of the survey procedures.

The cities surveys are conducted by an independent staff of interviewers recruited especially for that particular survey. These interviewers also undergo an extensive training program, which is essentially the same as that for the national sample, but with more time devoted to interviewing techniques. Interviewing is conducted over a 3-month period. A quality control program, consisting of supervisory observation and reinterview, is conducted similar to the national survey.

In general, the National Crime Panel has been well received by the respondents. The average response rate for the first 12 months of the national sample is about 96 percent. The response rate for the 13 cities interviewed thus far ranges from 92 percent to 98 percent.

The length of the interview varies, depending upon the number of household members interviewed as well as the number of victimizations in the household. Generally, the average time needed to complete the screen questionnaires for a household is approximately 20 minutes. Each incident report requires approximately 10 minutes for completion.

Periodically, supplemental inquiries may be added to the National Crime Panel. One such inquiry is a series of questions designed to collect data on a general attitude toward crime, the fear of crime, the effect of this fear on activity patterns such as choice of shopping area and places of entertainment, and the public's view of the police. There are two groups of questions--one group is asked only once in a household and refer to attitudes toward such things as respondent's neighborhood and shopping patterns. The second group of questions

are asked of each household member 16 years or older. These ask about individual attitudes toward such things as local police and fear of crime.

Data from the National Crime Panel will be compiled and published on a quarterly basis. Information will be produced on incidents that occurred during a particular quarter, rather than by victimizations that were collected during a particular quarter. Data for the July-September 1972 quarter will be tabulated from interviews conducted in August 1972 through March 1973.

These reports will provide crime rates by type of crime, victim characteristics, and geographic distribution. The types of crimes are not strictly defined according to definitions used in the Uniform Crime Report (UCR) issued by the Federal Bureau of Investigation, but in terms of the most serious aspect of the incident; i.e., the categories are descriptive of the incident itself. For example, in the survey, the two main subcategories of personal crimes are "Assaultive violence," which includes incidents with and without theft, and "Personal theft without assault." Robberies involving forcible attacks are classified as "Assaultive violence with theft," because the assault is the most serious aspect of the crime. Robberies involving only threats of harm, on the other hand, are classified under "Personal theft without assault." Using the strict UCR definitions, both of these incidents would be classified as "Robberies." However, classifications comparable to the UCR classifications can be obtained by combinations of several survey categories. For example, "serious assault, without theft" plus "attempted assault, with weapon" are comparable to the UCR definition of "aggravated assault."

Although the NCS is now operational and data are being produced on a regular basis, several methodological questions require further investigation to ensure collection of accurate and meaningful data on a continuing basis. The reinterview sample will be used to provide measures of the net difference rates and an index of inconsistency for various items. These are measures that reflect the reliability and validity of the original responses. Reinterview results are not available at this time.

An area of concern, described earlier, relates to self-response versus a household respondent, particularly for 12-and 13-year olds. Currently, data for persons 14 years old and over are obtained from each individual and for persons 12 and 13 years old from the household respondent, usually a parent. The primary reason for following this procedure concerns the sensitivity of the questions as they relate to the young children and their parents. Would parents allow their children to be interviewed? If they did allow it, would the children admit to all of their victimizations, particularly if the parents were present during the interview? How would the children themselves react to the questions and the interview situation itself? In January 1974, a test of self-response for this age group (the 12-and 13-year olds) will be conducted in San Francisco. Half the sample, about 5,000 households, will be interviewed using the self-response procedure for all persons 12 years old and over. In the other half, interviewers will use the current procedure of self-response for persons 14 years old and over, with a household respondent reporting for 12-and 13-year olds.

Methods of interview in the National Crime Panel should be tested further. Currently, the initial contact with the household is by personal visit, with telephone interview permitted for all sample persons 14 years old and over not present or available for interview at that time. Data collected from personal interviews will be compared to that collected by telephone to determine if there are any significant differences. In addition, a test should be designed to evaluate an initial telephone contact with a household to determine the effect on response rates, accuracy of reporting, and cost implications.

As part of the developmental work for this program, a test was designed to determine the feasibility of self-enumeration by mail. Screening questionnaires were mailed to approximately 5,200 households; approximately 2,800 (54 percent) households responded to the screening questionnaire. Of these, 555 households (19.8 percent) were identified as definitely having been victimized, while 491 households (17.5 percent) did not supply enough information to determine if an incident had occurred. These 1,046 households were all visited by an interviewer, either to complete the incident report or to determine if an incident did occur, as well as a sample of the mail nonresponses. At this point, the response results, as well as the comparison of data from the mail versus personal interviews need to be evaluated further. It appears, however, that the mail, self-enumeration method of screening is not satisfactory.

In addition to the operational oriented research activities just described, further record check studies should be undertaken to compare information provided by the respondent to the interviewer with that reported to the police. Work in this area should include further testing of the recall period and experimentation with methods to improve recall by probing questions.

There are other areas which suggest research activities. One that may be of interest involves the level of underreporting by respondents. Does this affect only particular subgroups of the population? More important, however, is the measurement of underreporting. How can you estimate this level and then adjust the data accordingly?

Another point of interest is the response variability to individual questions. This is affected by the respondent reaction to the wording of the questions; i.e., what does

"neighborhood" mean? What does "threat" mean? Some work should be done in this area also.

A continuing survey such as the National Crime Panel presents the opportunity for refinement and improvement of survey methodology relating to the specific subject area surveyed. Necessarily, the effort in establishing the continuing program has been great and has precluded testing all of the techniques that would have been desirable. However, as new insights are developed and new problems are discovered, plans will be made for further study as resources can be made available.

REFERENCES

(1) Dodge, Richard W. and Turner, Anthony G., Methodological Foundations for Establishing a National Survey of Victimization, Invited Paper, Social Statistics Section, American Statistical Association Annual Meeting, 1971.

(2) U.S. Bureau of the Census, "Victim Recall Pretest (Washington, D.C.)," Demographic Surveys Division, June 10, 1970 (Unpublished)

(3) U.S. Bureau of the Census, "Household Survey of Victims of Crime Second Pretest (Baltimore, Maryland)," Demographic Surveys Division, November 30, 1970 (Unpublished)

A SIMULTANEOUS EQUATIONS MODEL OF THE EDUCATIONAL PROCESS: THE COLEMAN DATA REVISITED WITH AN EMPHASIS UPON ACHIEVEMENT*

Anthony E. Boardman, Otto A. Davis, Carnegie-Mellon University,
and Peggy R. Sanday, University of Pennsylvania

1. Introduction

The report, Equality of Educational Opportunity [6], the EEOR, acted as a watershed for research into educational production functions. Virtually all of the voluminous research in this area chooses verbal achievement as the sole achievement measure. Very few papers examine other measures such as non-verbal, reading or mathematical achievement.

With a single exception, Levin [12], no model of the educational process allows for feedback effects from one variable to another. Studies have found, for example, that a pupil's self concept and belief in his ability to control the environment are extremely important predictors for pupil achievement. But, as Mosteller and Moynihan point out in On Equality of Educational Opportunity, OEOEO [14], "could not such feelings of control be essentially a feedback reaction from reality? Bright students who got good marks might well feel good about themselves." Thus a model of the educational process should postulate pupil achievement and control of the environment as endogenous variables.

Our paper has two main purposes. First, it examines verbal, non-verbal, reading, mathematical and general informational achievement. Second, it estimates the achievement equations of a simultaneous equations model of the educational process.[1] The analysis may allow us to make important statements about the factors affecting different types of achievement.[2]

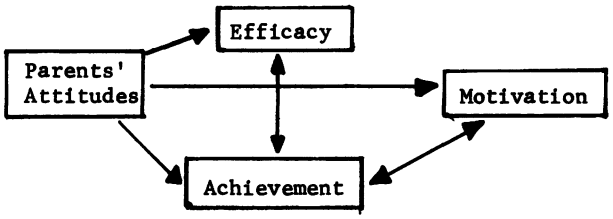
2. The Emphasis on Verbal Achievement

The EEOR [6] concentrated almost exclusively on verbal achievement.[3] Few reanalyses of the Equality of Educational Opportunity survey, EEOS, data consider any output other than verbal achievement. Mayeske, et al.[13], construct an index from the first component of a principal components analysis on verbal, non-verbal, reading, mathematical and general informational achievements. Boardman, et al. [3,4] derive a similar index. Most analyses consider only verbal achievement. In OEOEO [14], reanalyses by Jencks, Armor, Smith and Cohen, Pettigrew and Riley all use verbal achievement as the sole dependent variable.[4] Gordon [8] and Levin [12] also restrict attention to this achievement measure.

Many researchers have considered outputs other than verbal achievement. The list is too long to recite here, but Stafford [16], Aiken [1], and Dwyer [7] review many of them. One cannot really compare existing results or theories with this research for two reasons. First, previous research considers only a limited number of variables, sometimes only a single explanatory variable.[5] Second, prior studies do not use a simultaneous equations model.

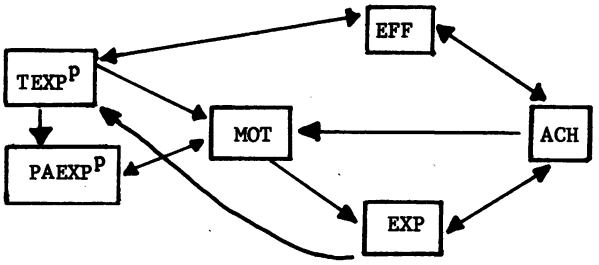
3. Simultaneous Equations Model of the Educational Process

Levin [12] should receive considerable credit for first publishing the notion of modeling the educational process by a system of simultaneous equations.[6] He estimated a model in which pupil achievement, motivation and efficacy, and parent's attitudes (expectations) interact thus:



More recently, Gordon [8] published a simply recursive model of the educational process with family structure, pupil's verbal ability, parental aspirations, and pupil's self-concept and aspirations as the endogenous variables. Because of Gordon's desire to use Path analysis rather than more sophisticated simultaneous equations techniques, the model does not allow any feedback effects. For this reason, Gordon's model represents a step backwards rather than a step forward from Levin's original formulation.

Boardman, et al. [3], extended Levin's work and successfully estimated a simultaneous equations model of the educational process with six endogenous variables. This model treats pupil achievement, ACH, motivation, MOT, expectations, EXP, and efficacy, EFF, and perceived parents' and teachers' expectations, $PAEXP^P$ and $TEXP^P$ [7], as endogenous variables. The following diagram represents the estimated relationships between the endogenous variables where the level of confidence exceeds 0.05 for all variables.



Of all the endogenous variables, only pupil efficacy and expectations appear to have a direct effect on pupil achievement; the other endogenous variables have important but indirect effects.

4. Description and Preliminary Analysis of the Achievement Tests

The Educational Testing Service, ETS, constructed the achievement tests and administered the questionnaires to the thousands of students in the EEOS. The verbal test consisted of thirty questions which asked for the "best" missing word of a sentence, and thirty questions on synonyms. The non-verbal test contained twenty-six questions on picking one figure from a group of five that had the least in common with the remaining four, and twenty-four questions on matching a given figure with one out of a group of five. The reading test required the students to read seven short passages (from articles, books, letters, sonnets or plays) and answer five questions per passage on content and tone. Twenty-five questions covered mathematics (simple computations and geometry). The last test consisted of ninety-five general informational questions that covered a wide range of interests and areas.[8] The ETS aimed to measure those "skills which are most important in our society for getting a good job and moving up to a better one, and for full participation in an increasingly technical world."[9] None of these tests were designed to measure intelligence.

Our first stage in the research consisted of performing a principal components analysis on the correlation matrix of the number of corrected answers to each test.[10] We obtained the following factor matrix:

Achievement Variable	Factor 1	Factor 2	Factor 3
Verbal	0.91447	-0.18853	0.14467
Non-verbal	0.79736	0.06406	-0.60147
Reading	0.88104	-0.29293	0.07729
Mathematical	0.77779	0.59325	0.17393
General Informational	0.89636	-0.09262	0.15909

Factor	Eigenvalue	Pct. of Var.
1	3.65667	73.1
2	0.48598	9.7
3	0.44422	8.9

Table I

The first component indicates that verbal achievement has most in common with the other achievement measures, while non-verbal achievement and mathematical achievement have least in common with the other achievement measures. The second and third components suggests that non-verbal and mathematical achievements have little in common with each other. This finding surprised us. In fact, both non-verbal and mathematical achievement correlate least with each other. The rapidly falling eigenvalues show that the first the first component explains most of the combined variance, while the other components add little. Basically, these tests measure a similar characteristic.[11]

5. Regional, Racial and Individual Findings for the Achievement Equations

In view of the above conclusion that the various tests probably measure the same characteristic, it is not at all surprising that the estimated results (reported in Appendix II[12] at the end of the paper) indicate that in general the same endogenous and exogenous variables explain each of the various tests. For example, efficacy, the endogenous variable which directly affects achievement, has positive and significant coefficients in the structural equations of all of the tests. Similarly, the coefficients for the average teachers' score are always positive, while those for the age of the student are always negative. Such general results may comfort those who have analyzed only one achievement measure.

The significance of the efficacy variable suggests that performance on all of these tests improves as the child increases his self-concept and belief in his ability to control the environment. These attitudes appear particularly important for general informational achievement. Of all the other endogenous variables, pupil expectations is the only one which enters the second stage achievement equations; it enters only the mathematical equation. The other endogenous variables, including motivation--a measure of hard work and attitude to work--fail to exert a direct effect.[13]

The coefficients for the dummy variables for the regions of the U. S. vary slightly across regions. The variables in the non-verbal achievement equation seem quite different from those in the verbal achievement equation, yet quite similar to those in the general informational equation. Some consistencies emerge clearly. Students from the Plains States seem to perform better than students from any other region while students from the South, both the Southeast and the Southwest, appear to do worse than students from the other regions. Perhaps the most striking finding is that these coefficients are relatively small in absolute value, while the difference between regional mean achievements are quite substantial.

Substantial differences exist in the average achievement scores across the ethnic groups (see Table II). American Indians, Mexican Americans, Blacks, and Puerto Ricans obtain on the average 12 to 14 fewer correct answers than Whites on the verbal achievement test. Oriental-Americans obtain on the average 2 fewer correct answers than Whites on this test. When we take other variables into account, by including them

in the regressions, the differences drops substantially. The structural form coefficients for Blacks and Whites differ by approximately 5 points in the verbal achievement equation, a drop of 9 points. [14] Similar patterns hold for the other minority groups except for Oriental Americans who have more positive structural form coefficients than Whites. The coefficients for American Indians and Whites differ by approximately 4 in the verbal achievement equation, a reduction of about 8 points. For Mexican Americans and Puerto Ricans the initial differences reduce to approximately 5 points. Hence while minority group status appears to be detrimental for four of these groups, the differentials are not nearly so substantial as might be suggested by a simple examination of the averages.

Pupils attending predominantly White schools (70%-100% White) perform better than pupils in partially integrated schools (30%-69% White) or mainly Black schools (0%-29% White). Except for verbal achievement, there appears to be negligible benefits in achievement from attending an integrated school as opposed to a non-majority school. These results suggest that if one wants to integrate to improve achievement, the integration should be complete. [15]

Average socio-economic class of peers is positive in all equations. This variable may reflect a peer group orientation to achievement. Eliot Richardson [16] said that children learn more from each other than from any other resource of the education environment. If this is the case then the values communicated among peers could have an important impact on a child's receptivity to learning. Average SES of the school could also reflect the general quality of the school, or something about the home background. When this variable is excluded from the regressions the school variables change more than the home variables. Hence one might infer that it reflects the school more than the home.

One reason for including the pupils' average socio-economic status stems from the criticism by educators and sociologists that one cannot reasonably consider teacher and school effects as exogenous with individual pupil data[17]. The argument claims that better pupils attract better teachers. Furthermore, those pupils of a higher socio-economic status may attend better schools because their parents can afford (may be required) to pay more per pupil to the school board. Thus, both the quality of the teachers and the schools may be superior in a higher socio-economic area. If one finds that school and teacher variables are important, it may be a result of better pupils, not better schools. Since this research controls for the average socio-economic status any observed teacher and school effects should not be spurious.[18]

The variable for sex is interesting. For verbal, non-verbal and mathematical achievement, as well as for general knowledge, the estimated coefficients are negative and very significant. These results indicate that males are better achievers across these individual cognitive dimensions. On the other hand, in the test for reading achievement, the estimated coefficient for sex is positive and significant, which indicates that on the average females are better at reading than are males. Apparently this phenomenon has been observed many time previously, and some sociologists and psychologists have attempted to explain it by saying that our society considers reading to be a feminine rather than a masculine activity. [19]

In regard to other individual characteristics, observe that age has a negative effect upon all measures of achievement. One may expect that schools hold back some underachieving students. In the twelfth grade, these pupils would be older yet still poorer performers. The more older brothers and sisters that a pupil has, the worse he does on all measures of achievment, with the exception of mathematical achievment. Interestingly home stability as measured by whether there are two parents alive and living at home seems important for non-verbal and mathematical achievements. Information in the home seems important for verbal and general informational achievements.

6. School Variables Which May Affect Achievement

Recent years have witnessed an increasing acceptance of the argument that variables associated with schools contribute little to educational outcomes.[20] Our results do not support this position. Even though our measures of school characteristics are crude and certainly not ideal, they do appear to have important effects on achievement. The best measure that can be obtained for the quality of a school's faculty, for example, is the average score of the teachers on a verbal achievement test. The coefficients for teachers' average verbal right in the structural equations for each achievement test is positive and exceptionally significant. Similarly, the number of teachers per pupil, often thought in some educational circles (but not among laymen) to be an irrelevant variable, is positively and significantly associated with each of the various measures of achievement.

Teachers' experience, measured by the aveage number of years teaching, appears to have a quadratic effect upon all measures of achievement except mathematical achievement. A simple interpretation of this effect is that in the first few years on the job, a teacher loses the initial excitement and enthusiasm and thus performs less well; but as years pass, experience begins to dominate and has an increasingly

positive effect upon achievement. One might also argue that natural selection occurs and dedicated teachers tend to be the ones who remain on the job to gain experience while those who really were not interested in this profession drop out.

The above results are highlighted and perhaps confounded by the fact that the number of teachers leaving is positively associated with achievement as it is measured on the verbal, non-verbal and reading tests. Also somewhat surprisingly, the perception on the part of teachers of the lack of effective administrative leadership is positively related to all measures of achievement. Since the mean of this variable is low, one might speculate that only the better and more perceptive teachers are able to recognize such problems and these teachers perform well in any event.

Schools which have a policy of administering achievement and IQ tests to their students also have pupils who score significantly higher on each of the various achievement tests. Even school facilities, generally thought to be irrelevant, appear to be positively associated with non-verbal and reading achievement. The age of the school is negatively associated with verbal and non-verbal achievement, but positively associated with reading achievement. Finally, problems in the school have negative effects on all achievement measures.

Unfortunately this body of data does not include variables which measure the degree of interaction between pupils and teachers in the classroom, nor does it include measures of teaching materials. In retrospect, we believe that we should have included a variable for the curriculum program. One rarely included all important variables in an estimation. We aim to perform further analyses on these rich data in later papers.

7. Concluding Remarks

These results do not allow us to say directly that the school is more important than the home for one type of achievement, but not for another type. Both the home and the school are important for all achievements, especially verbal and general informational. More variables seem important for non-verbal achievement than for any other type of achievement. The absolute value of the coefficients in the mathematical achievement equations are generally smaller than in the other equations. This finding and the lower R^2 indicated that the explanatory variables may be less important for mathematics than for other achievements. Perhaps mathematics requires a specific attitude or aptitude more than other subjects require a distinct attitude or aptitude. Contrary to the probably expectations of the EEOR's authors, the general informational equation fits the data best, not the verbal equation.

There are several other conclusions which require emphasis. First, of course, is the conclusion that the various tests really measure a common characteristic. Furthermore, an independent variable which affects one measure of achievement generally affects the others in the same direction and in roughly the same magnitude. This finding should offer comfort to those who have just used one measure of achievement in their analysis.

Relative to the omitted group (the surprisingly large number of American students who state that they do not know their race), Whites and Orientals perform best on all tests of achievement. Nevertheless, the twelve to fourteen mark differential between the other minority groups and Whites in verbal achievement narrows to four or five marks when all other factors are controlled by inclusion.

Quite substantially these results show that good teachers and good schools are important for educational achievement. Teachers average verbal right, class size, teachers' experience, school facilities and problems in the school have significant and important effects on the achievement measures. These variables are important components in the educational process.

Table II

Achievement Test

	Ver.	Non-V.	Reading	Math	Gen.Info.
B.	23.13	27.37	16.83	6.68	38.81
	(11.37)	(8.58)	(6.42)	(3.17)	(11.55)
W.	37.05	36.16	23.24	10.97	54.18
	(12.18)	(7.00)	(6.21)	(4.61)	(12.52)
P.R.	23.75	28.40	16.62	6.85	37.66
	(11.92)	(9.19)	(6.77)	(3.50)	(13.17)
M.A.	24.32	28.97	16.89	7.70	40.07
	(11.77)	(9.42)	(6.71)	(3.53)	(12.78)
Or.	34.67	36.48	21.45	10.99	50.78
	(13.24)	(7.88)	(6.62)	(4.75)	(12.60)
A.I.	24.56	31.51	17.39	7.83	42.41
	(12.17)	(8.25)	(6.44)	(3.67)	(13.04)
O.	29.19	32.15	18.56	8.56	44.71
	(13.46)	(8.68)	(7.23)	(4.20)	(14.39)

Average number of correct responses on the achie achievement tests across races--(standard deviations in parentheses)

B.= Black, W.= White, P.R.= Puerto Rican,
M.A.= Mexican American, Or.= Oriental,
A.I.= American Indian, O.= Other

APPENDIX I: DESCRIPTION OF THE VARIABLES

Abbreviation	Description	Mean	Standard Deviation
VR	Verbal Achievement	28.654	13.497
NVR	Non-verbal Achievement	31.512	9.107
RR	Reading Achievement	19.124	7.100
MR	Mathematical Achievement	8.597	4.347
GITR	General Informational Achievement	44.796	14.214
ACH	Achievement	0.099	3.664
MOT	Motivation	0.006	2.041
EXP	Expectations	0.020	1.666
EFF	Efficacy	0.007	3.271
$PAEXP^P$	Perceived Parents' Expectations	0.018	2.332
$TEXP^P$	Perceived Teachers' Expectations	-4.269	1.615
CONST	Constant	1.000	0.000
NEWENG	New England	0.028	0.165
MIDATL	Mid-Atlantic	0.215	0.411
LAKES	Great Lakes	0.149	0.356
PLAINS	Plains	0.045	0.206
SEAST	Southeast	0.215	0.411
SWEST	Southwest	0.097	0.295
BLACK	Black	0.265	0.441
WHITE	White	0.275	0.447
PRICAN	Puerto Rican	0.082	0.275
MEXAM	Mexican American	0.147	0.354
ORIENT	Oriental	0.081	0.273
AMIND	American Indian	0.081	0.273
PWPICLY	Proportion of white pupils in class last year	3.135	1.477
MLYBLCK	Mainly black school	0.366	0.482
MIX	Integrated school	0.101	0.302
SES	Socio-economic status	0.080	2.307
AVSES	Average socio-economic status	0.080	1.099
INFO	Information available	0.051	1.763
SMSA	Metropolitan Area	1.332	0.471
SEX	Sex	3.010	0.998
AGE	Age	4.067	0.916
NOBAS	Number of older brothers and sisters	2.877	2.159
TWOP	Two parents	0.642	0.479
FL	Foreign Language	3.219	1.071
RBS	Reading before school	2.395	1.199
PTAS	Parents talking about school	2.009	1.117
PTAAT	Parents attend PTA	1.702	1.024
NHWTV	Watching television	3.969	2.119
NHWTV2	(Watching TV)^2	20.244	16.901
TC	This city	0.755	0.430
NTCHSCL	Number of times changed school	2.586	1.524
LSTCHSCL	Last time changed school	6.004	1.651
TAVR	Teachers' average verbal right	24.382	2.295
NTPRPUP	Number of teachers per pupil	0.044	0.008
TANYTCH	Teachers' average number of years teaching	4.430	0.693
TANYTCH2	(Teachers' average number of years teaching)^2	20.108	6.196
PWTCHLY	Proportion of white teachers in class last year	3.647	1.626
TASEX	Teachers' sex	2.924	0.283
TPTC	Proportion of teachers from this city	0.426	0.255
PROBLEMS	Problems in the school	167.75	2.389
FACILITS	School facilities	12.346	1.799
AGES	Age of school	4.778	1.757
NTCHLV	Number of teachers who leave	2.152	1.396
TPADTN	Teachers' problems with administration	0.114	0.146
PRNMADEG	Principal has Master's degree	4.213	0.642
TEST	Testing experience	1.710	0.485
NTLKGC	Number of times talk to guidance counsellor last year	2.531	1.262

APPENDIX II: REDUCED FORM AND STRUCTURAL FORM ESTIMATES OF THE ACHIEVEMENT EQUATIONS

Dependent Variable	Verbal Achievement		Non-verbal Achievement		Reading Achievement		Math Ach
Explanatory Variable	Reduced Form	Structural Form	Reduced Form	Structural Form	Reduced Form	Structural Form	Reduced Form
ACH							
MOT							
EXP							
EFF		1.595 (12.792)		0.889 (9.944)		0.915 (13.450)	
PAEXP^P							
TEXP^P							
CONST	45.924 (6.172)	51.918 (7.288)	42.225 (8.074)	44.535 (8.750)	18.799 (4.576)	21.552 (5.965)	17.089 (6.724)
NEWENG	0.283 (0.481)		-1.515 (-3.671)	-1.333 (-3.568)	0.817 (2.521)	0.575 (2.023)	0.198 (0.988)
MIDATL	0.678 (1.904)		-0.798 (-3.188)	-0.927 (-5.073)	0.508 (2.583)		0.033 (0.273)
LAKES	0.240 (0.755)		-0.091 (-0.408)		0.245 (1.398)		0.156 (1.432)
PLAINS	1.990 (4.293)	0.750 (1.769)	0.391 (1.200)		1.760 (6.876)	0.951 (4.130)	0.857 (5.410)
SEAST	0.549 (1.439)		-1.192 (-4.450)	-1.489 (-7.162)	0.717 (3.407)		0.107 (0.820)
SWEST	-0.671 (-1.778)	-1.276 (-4.223)	0.157 (0.593)		0.418 (2.007)		-0.023 (-0.178)
BLACK	-1.508 (-3.790)	-2.642 (-6.882)	-1.302 (-4.658)	-2.207 (-10.137)	0.078 (0.357)	-0.360 (-2.182)	-0.986 (-7.259)
WHITE	4.115 (10.354)	2.408 (5.925)	2.222 (7.956)	1.085 (4.758)	2.485 (11.323)	1.827 (10.355)	1.354 (9.972)
PRICAN	-2.141 (-4.573)	-2.109 (-4.678)	-1.762 (-5.356)	-1.988 (-7.182)	-0.733 (-2.835)	-0.460 (-2.194)	-0.707 (-4.424)
MEXAM	-2.592 (-6.337)	-2.961 (-7.493)	-1.964 (-6.832)	-2.344 (-10.647)	0.830 (-3.679)	-0.792 (-4.607)	-0.643 (-4.600)
ORIENT	4.131 (8.941)	3.811 (8.632)	3.033 (9.340)	2.706 (9.859)	2.130 (8.353)	2.119 (10.028)	1.961 (12.425)
AMIND	-1.664 (-3.576)	-1.829 (-4.094)	0.538 (1.644)		-0.255 (-0.992)		-0.432 (-2.717)
PWPICLY	-0.008 (-0.077)		0.033 (0.437)		0.060 (1.028)		-0.028 (-0.774)
MLYBLCK	-2.448 (-7.926)	-2.211 (-8.370)	-0.277 (-1.274)		-0.647 (-3.798)	-0.677 (-4.459)	-0.698 (-6.621)
MIX	-1.609 (-4.628)	-1.770 (-5.597)	-0.602 (-2.464)	-0.643 (-3.018)	-0.493 (-2.567)	-0.723 (-4.105)	-0.462 (-3.894)
SES	0.972 (19.405)	0.717 (13.813)	0.394 (11.182)	0.246 (6.698)	0.368 (13.315)	0.225 (7.974)	0.230 (13.438)
AVSES	1.788 (15.267)	1.718 (15.584)	0.757 (9.198)	0.717 (9.362)	0.640 (9.894)	0.588 (10.388)	0.321 (8.025)
INFO	0.404 (6.640)	0.120 (1.898)	0.201 (4.694)		0.165 (4.925)		0.128 (6.146)
SMSA	0.194 (0.776)	-0.629 (-2.693)	0.016 (0.093)	-0.445 (-2.673)	0.417 (3.019)		0.259 (3.034)
SEX	-0.204 (-2.356)	-0.568 (-6.410)	-0.384 (-6.316)	-0.607 (-9.623)	0.535 (11.204)	0.336 (6.953)	-0.661 (-22.375)
AGE	-0.945 (-9.661)	-0.498 (-4.915)	-0.688 (-10.004)	-0.446 (-6.248)	-0.699 (-12.946)	-0.461 (-8.378)	-0.229 (-6.866)

Dependent Variable	Verbal Achievement		Non-verbal Achievement		Reading Achievement		Math Ach
Explanatory Variable	Reduced Form	Structural Form	Reduced Form	Structural Form	Reduced Form	Structural Form	Reduced Form
NOBAS	-0.355	-0.233	-0.220	-0.153	-0.209	-0.148	-0.037
	(-8.461)	(-5.530)	(-7.445)	(-5.150)	(-9.031)	(-6.505)	(-2.566)
TWOP	0.314		0.617	0.503	0.242		0.167
	(1.663)		(4.645)	(3.871)	(2.320)		(2.591)
FL	-0.618	-0.567	-0.367	-0.342	-0.199	-0.155	-0.080
	(-6.783)	(-6.444)	(-5.738)	(-5.487)	(-3.964)	(-3.223)	(-2.586)
RBS	0.450		0.056		0.233		-0.004
	(5.757)		(1.026)		(5.402)		(-0.143)
PTAS	0.581		0.316		0.176		0.128
	(6.996)		(5.414)		(3.837)		(4.497)
PTAAT	-0.446		-0.240		-0.332		-0.030
	(-4.991)		(-3.812)		(-6.717)		(-0.977)
NHWTV	0.446	-0.199	0.563	0.301	0.562	0.298	0.135
	(2.352)	(-4.935)	(4.226)	(2.248)	(5.367)	(2.888)	(2.078)
NHWTV2	-0.095		-0.076	-0.035	-0.082	-0.040	-0.033
	(-3.986)		(-4.509)	(-2.043)	(-6.188)	(-3.044)	(-4.012)
TC	-0.227		0.647	0.590	0.356	0.284	-0.113
	(-1.051)		(4.257)	(3.967)	(2.982)	(2.524)	(-1.531)
NTCHSCL	0.023		0.167	0.202	0.019		-0.057
	(0.354)		(3.631)	(4.489)	(0.535)		(-2.549)
LSTCHSCL	0.293		0.300	0.192	0.198	0.078	0.120
	(4.969)		(7.249)	(4.526)	(6.091)	(2.535)	(5.980)
TAVR	0.575	0.499	0.608	0.619	0.364	0.323	0.091
	(8.798)	(9.480)	(13.228)	(14.982)	(10.085)	(10.273)	(4.058)
NTPRPUP	23.076	36.405	19.935	23.697	21.033	25.231	2.534
	(1.882)	(3.346)	(2.314)	(2.894)	(3.108)	(4.225)	(0.605)
TANYTCH	-5.998	-4.404	-3.944	-2.804	-2.580	-1.627	-0.843
	(-5.322)	(-4.081)	(-4.979)	(-3.667)	(-4.147)	(-2.779)	(-2.190)
TANYTCH2	0.720	0.495	0.473	0.324	0.323	0.189	0.119
	(5.685)	(4.092)	(5.315)	(3.768)	(4.617)	(2.893)	2.740
PWTCHLY	0.038		0.253	0.209	0.164	0.097	0.006
	(0.415)		(3.886)	(4.023)	(3.213)	(2.315)	(0.174)
TASEX	-0.920	-0.580	-0.939	-0.799	-0.296		-0.296
	(-2.498)	(-1.771)	(-3.626)	(-3.217)	(-1.455)		(-2.349)
TPTC	-0.046		-0.023		-0.199		-0.296
	(-0.087)		(-0.062)		(-0.688)		(-1.658)
PROBLEMS	-0.082	-0.113	-0.090	-0.109	-0.037	-0.057	-0.036
	(-2.085)	(-2.975)	(-3.275)	(-4.083)	(-1.698)	(-2.826)	(-2.698)
FACILITS	-0.060		0.005		0.038	0.047	-0.051
	(-1.102)		(0.136)		(1.270)	(1.657)	(-2.758)
AGES	-0.097		-0.118	-0.123	0.058	0.078	-0.018
	(-1.607)		(-2.789)	(-3.138)	(1.750)	(2.561)	(-0.864)
NTCHLV	0.192	0.139	0.137	0.117	0.104	0.068	0.045
	(2.773)	(2.115)	(2.826)	(2.479)	(2.729)	(1.911)	(1.887)
TPADTN	3.719	3.188	1.527	1.207	1.501	1.153	0.504
	(5.920)	(5.373)	(3.458)	(2.836)	(4.328)	(3.505)	(2.349)
PRNMADEG	-0.728		-0.391		-0.577		-0.171
	(-5.054)		(-3.864)		(-7.253)		(-3.473)
TEST	1.109	0.984	0.499	0.463	0.314	0.250	0.445
	(5.341)	(5.077)	(3.417)	(3.333)	(2.743)	(2.389)	(6.281)
NTLKGC	0.712		0.485		0.384		0.215
	(9.960)		(9.656)		(9.722)		(8.810)
MLR^2	0.3606	0.4330	0.3063	0.3590	0.2958	0.3897	0.2809
$ALTR^2$	0.3606	0.3560	0.3063	0.3028	0.2958	0.2908	0.2809

Dependent Variable	Math Ach	General Informational Achievement	
Explanatory Variable	Structural Form	Reduced Form	Structural Form
ACH			
MOT			
EXP	0.314 (3.017)		
EFF	0.354 (5.755)		2.124 (16.344)
$PAEXP^P$			
$TEXP^P$			
CONST	16.454 (7.282)	68.931 (8.937)	74.780 (10.406)
NEWENG		-0.005 (-0.009)	
MIDATL		-0.126 (-0.341)	-0.935 (-3.133)
LAKES	0.185 (2.151)	0.214 (0.649)	
PLAINS	0.595 (4.038)	2.183 (4.544)	
SEAST		-0.066 (-0.167)	-1.217 (-3.989)
SWEST	-0.178 (-1.724)	-0.319 (-0.814)	-1.284 (-3.796)
BLACK	-1.414 (-10.957)	-1.512 (-3.667)	-2.610 (-8.227)
WHITE	0.909 (6.285)	4.756 (11.545)	3.116 (8.943)
PRICAN	-0.741 (-4.822)	-3.229 (-6.656)	-2.730 (-6.498)
MEXAM	-0.698 (-5.177)	-3.023 (-7.132)	-2.928 (-8.757)
ORIENT	1.894 (12.259)	4.393 (9.174)	4.113 (9.913)
AMIND	-0.496 (-3.245)	-0.458 (-0.949)	
PWPICLY		0.102 (0.925)	
MLYBLCK	-0.586 (-6.408)	-2.204 (-6.885)	-2.188 (-7.972)
MIX	-0.560 (-5.091)	-1.350 (-3.747)	-1.434 (-4.224)
SES	0.118 (5.151)	0.842 (16.205)	0.525 (9.503)
AVSES	0.261 (7.270)	1.594 (13.134)	1.491 (13.733)
INFO		0.560 (8.882)	0.227 (3.414)
SMSA		0.883 (3.407)	
SEX	-0.732 (-21.071)	-1.803 (-20.097)	-2.267 (-24.338)
AGE	-0.829 (-2.434)	-0.822 (-8.108)	-0.260 (-2.442)

Dependent Variable	Math Ach	General Informational Achievement	
Explanatory Variable	Structural Form	Reduced Form	Structural Form
NOBAS		-0.455 (-10.465)	-0.310 (-6.981)
TWOP	0.112 (1.809)	0.342 (1.745)	
FL		-0.267 (-2.832)	-0.168 (-1.797)
RBS		0.599 (7.404)	
PTAS		0.456 (5.295)	
PTAAT		-0.497 (-5.361)	
NHWTV		0.591 (3.004)	
NHWTV2		-0.100 (-4.055)	
TC		0.260 (1.162)	
NTCHSCL		0.006 (0.092)	
LSTCHSCL	0.071 (3.788)	0.391 (6.407)	0.122 (2.108)
TAVR	0.072 (3.960)	0.565 (8.337)	0.506 (8.199)
NTPRPUP	7.064 (1.914)	30.320 (2.386)	43.785 (3.614)
TANYTCH		-7.167 (-6.134)	-5.088 (-4.467)
TANYTCH2		0.867 (6.607)	0.566 (4.396)
PWTCHLY		0.292 (3.039)	
TASEX	-0.205 (-1.834)	-0.663 (-1.736)	
TPTC	-0.257 (-1.928)	-1.315 (-2.426)	-0.934 (-1.886)
PROBLEMS	-0.045 (-3.571)	-0.110 (-2.708)	-0.145 (-3.675)
FACILITS		-0.115 (-2.044)	
AGES		-0.050 (-0.797)	
NTCHLV		0.136 (1.892)	
TPADTN	0.442 (2.175)	2.742 (4.211)	2.326 (3.684)
PRNMADEG		-0.604 (-4.045)	
TEST	0.447 (6.596)	0.901 (4.187)	0.839 (3.953)
NTLKGC		0.812 (10.950)	
MLR^2	0.3362	0.3806	0.4745
$ALTR^2$	0.2750	0.3806	0.3772

Bibliography

[1] Aiken, L.R., "Ability and Creativity in Mathematics," Review of Educational Research, 43, Fall 1973, pp. 405-432.

[2] Boardman, A.E., Review of On Equality of Educational Opportunity, Mosteller and Moynihan (eds.), Journal of the American Statistical Association, 68 June 1973, pp. 489-491.

[3] Boardman, A.E., Davis, O.A. and Sanday, P. R., "A Simultaneous Equations Model of the Educational Process," School of Urban and Public Affairs, Carnegie-Mellon University (submitted for publication).

[4] Boardman, A.E., Davis, O.A. and Sanday, P. R., "Education From An Anthropological Perspective: An Empirical Investigation of Structural Differences Among Blacks and Whites," School of Urban and Public Affairs, Carnegie-Mellon University, presented at the Conference on the Contributions of Anthropology to Public Policy Formulation, Philadelphia, October 1973.

[5] Buros, O.K., (ed.), Fourth and Fifth Measurements Yearbook, Gryphon Press, Highland Park, N.J., 1953 and 1959.

[6] Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., and R.L. York, Equality of Educational Opportunity, Washington, D.C., U.S. Government Printing Office, 1969.

[7] Dwyer, C.A., "Sex Differences in Reading: An Evaluation and a Critique of Current Theories," Review of Educational Research, 43, Fall 1973, pp. 455-467.

[8] Gordon, C., "Looking Ahead: Self-conceptions, Race and Family as Determinants of Adolescent Orientation to Achievement," Rose Monograph Series, American Sociological Association, 1973.

[9] Hanushek, E.A., Education and Race, Lexington: D.C. Heath, 1972.

[10] Hechinger, F.M., "Home is a Crucial Factor," New York Times, Section E., p.9, May 27, 1973.

[11] Jencks, C., et al., Inequality, Basic Books, 1972.

[12] Levin, H.M., "A New Model of School Effectiveness," in Do Teachers Make a Difference?, U.S. Government Printing Office, 1970, pp.25-78.

[13] Mayeske, G.W., et al., "A Study of our Nation's Schools," working paper, U.S. Department of Health, Education and Welfare, Office of Education (not dated).

[14] Mosteller, F. and Moynihan, D.P. (eds.), On Equality of Educational Opportunity (OEOEO), Random House, 1972.

[15] Purves, A. C., Literature Education in Ten Countries, Wiley: New York, 1973.

[16] Stafford R.E., "Hereditary and Environmental Components of Quantitative Reasoning," Review of Educational Research, 42, Spring 1972, pp. 183-201.

[17] Toward Equal Educational Opportunity. The report of the Select Committee on equal educational opportunity, United States Senate, U.S. Government Printing Office, Washington, D.C., 1972.

FOOTNOTES

* The authors thank Professors Timothy McGuire, Joseph Kadane and Edwin Fenton of Carnegie-Mellon University, and Professor Henry Levin of Stanford University for helpful comments on earlier drafts of this and related work. David Rattner, now of Princeton University, performed invaluable programming assistance during the past summer. Finally, we are indebted to the Ford Foundation, the U. S. Office of Education, and the National Science Foundation for financial assistance. The authors accept full responsibility for the opinions expressed in this paper and for any remaining errors.

1. We estimated the model by two stage least squares. Our sample consisted of over 16,000 twelfth grade students from all regions of the country and with different ethnic backgrounds.

2. Several such hypotheses can be found in the first reports of the on-going study by the International Association for the Evaluation of Educational Achievement, IEA. See Hechinger[10].

3. Ambiguously, the Coleman Report refers to some tests as measures of ability, and some as measures of achievement. We prefer to regard them all as achievement measures.

4. See Boardman [2], for a thorough review of OEOEO.

5. See Dwyer [7], for example.

6. For a thorough view of Levin's work, see Boardman, et al. [3]

7. Appendix I contains brief operational definitions of these variables. More detailed descriptions are available upon request.

8. The ETS took the verbal test from the School and College Ability Tests, SCAT. The non-verbal test came from the Interamerican Tests of General Ability. The reading and mathematical tests were each one-half of a test from the Sequential Tests for Educational Progress, STEP. The ETS based the general informational test questions on items used in their earlier research studies. These comments apply only to the ninth and twelfth grade tests. More information on some of these tests appear in the Mental Measurements Yearbooks [5].

9. See the EEOR, p. 20.

10. The ETS calculated a scale score for the verbal, non-verbal and reading tests, but not for the other tests. We could have corrected for guessing, but the instructions specifically stated that the students' score depended on the number of correct answers.

11. The above results suggested that there was only a single latent factor. We performed a factor analysis with squared multiple correlations as communality estimates and found strong evidence of only one factor. A varimax rotation on the factor matrix for the cases N=2 and N=3 suggested that even if the second factor was not in error, it was not a non-verbal factor (on the varimax rotated factor matrix for N=2 mathematical right had the highest loading of 0.73 followed, in order of magnitude, by general informational right with a loading of 0.52).

12. All variables in the structural equations have a level of confidence in excess of 0.95 for a one tailed test. The table presents t-statistics in parentheses. MLR^2 means the R^2 is calculated using observed values of endogenous variables; $ALTR^2$ uses predicted values.

13. Single equation estimation techniques are likely to show that motivation has a sifnificant direct effect. For example, see Hechinger's article [10], on the recent IEA findings.

14. The research classifies students who do not consider themselves members of the given racial groups as "Other"; we excluded this category from the regressions.

15. To answer this question more fully, one should consider the ethnic groups individually. See Boardman et al. [4]

16. See Toward Equal Educational Opportunity [17] p. 235.

17. See, for example, Jencks in OEOEO [4], pp. 82-83.

18. Average socio-economic status acts like the IEA's sailing handicap. See, for example, Purves [15] pp. 121-125.

19. See Dwyer [7] for a full discussion of the alternative theories.

20. See Jencks [11], for example.

AN ECONOMETRIC MODEL FOR ESTIMATING IQ SCORES AND ENVIRONMENTAL INFLUENCES ON THE PATTERN OF IQ SCORES OVER TIME*

Joseph B. Kadane, Carnegie-Mellon University
Timothy W. McGuire, Carnegie-Mellon University
Peggy R. Sanday, University of Pennsylvania
Richard Staelin, Carnegie-Mellon University

1. INTRODUCTION

We offer in this paper a preliminary analysis of the effects of a semi-segregated school system on the IQ's of its students. We offer it with educational policy objectives in mind. Our basic data consist of IQ scores for a panel of children at kindergarten, fourth, sixth, and eighth grades and associated environmental data obtained from their school records. We developed a statistical model to analyze longitudinal data when both process error and measurement error must be accounted for. Our statistical model can be used on longitudinal data with other measures than IQ.

We are aware of confusion about just what IQ is, or, put another way, whether IQ is anything but what an IQ test measures. While we use IQ tests in this paper, we use them as convenient measures of a certain kind of performance thought to be important for success in schools and certain kinds of jobs. Sanday (1972 a,b,c; 1973) gives a critique of IQ tests. She says that "the content of test items is often related to experience and learning which only middle and upper class children would be likely to be exposed to" (Sanday 1972 a: 420). This suggests that the nature and degree of contact with mainstream culture would have an impact on IQ scores. We interpret our results with this theory in mind. Most of the environmental variables included in our model can be construed to measure the nature or degree of contact with mainstream culture.

2. STATISTICAL MODEL

In structuring our model, we quickly found that we had to distinguish two different phenomena, measurement error (different measures of IQ of the same person on successive days) and process error(individual variability from our notion of how IQ's develop and change over time).

We begin with measurement error. Let Z_i^j be the i-th student's test IQ score at grade j (j=1 for kindergarten, 2 for 4th grade, 3 for 6th grade, 4 for 8th grade), and X_i^j be the i-th student's true (but unobservable) IQ score at grade j. Then

$$Z_i^j = X_i^j + u_i^j , \tag{1}$$

where u_i^j is the measurement error. We made the standard assumptions for u_i^j: i.e., $E(u_i^j)=0$, $E(u_i^j)^2 = \tau^j$, $E(u_i^j u_i^{j'}) = E(u_i^j u_i^{j'}) = 0$, u_i^j distributed multivariate normal. In other words the measurement error for each student has a variance τ^j, which causes the test scores to differ from the true score but is uncorrelated with any other test score and is independent of the student's true score. (This concept is usually referred to as the standard error of measurement, and $\tau^j = (15)^2(1-r) \approx 25$, where r is the reliability.) Equation (1) is rather firmly rooted in our notion of what measurement error is. Notice that stopping here gives a model with more parameters than data.

The next step in our model specification is to state how we think the "true IQ's", X_i^j, change over time in response to the environment and changes in it. This is done in the following equations:

$$\begin{aligned} X_i^j &\sim N(\underline{W}_i^j \underline{\beta}^j, \sigma^j), \qquad j=1, \\ X_i^j &\sim N(X_i^{j-1} + \underline{W}_i^j \underline{\beta}^j, \sigma^j), \quad j=2,3,4, \end{aligned} \tag{2}$$

where $\underline{W}_i^j$ is a vector of demographic and environmental variables such as race, sex, SES of peers, etc. (discussed in section 4) and $\underline{\beta}^j$ is a vector of weights. In other words, the student's true score centers around the previous true IQ (except at kindergarten) modified by the effects of demographic and environmental factors.

What we mean by the above is that at kindergarten (before the test) we have no hard information about the child's true IQ score. Thus we express our beliefs in the form of a distribution (normal) with a mean (based on the demographics) and a variance. For the other years we also have opinions on the child's true IQ scores. These center around his previous true unobserved score plus the effect of contacts with the environment since our last estimate.

Equations (2) are our priors about the true but unobservable IQ scores for each student. These priors involve parameters (technically called hyperparameters) which can be estimated (see section 3) from the observed data, namely the four IQ scores and the vector of demographic and environmental variables.

We are not completely diffuse about our knowledge of some of the parameters of the prior. In particular, we believe that the variance σ^1 is quite large (i.e., about 200) since we have little relevant information about the child before he or she enters the school system. The variances σ^2, σ^3, and σ^4, however, should be much smaller (i.e., somewhere near 20-30), since we have a) at least one observed test score and b) measures on a number of variables which might influence changes from the previous true score. The model as stated acknowledges our uncertainty about the unobservable true IQ score by defining X_i^j as a random variable which is completely described (in probability density terms) only after the $\underset{\sim}{\beta}$'s and $\underset{\sim}{\sigma}$'s are known.

The parameter space for the model specified by (1) and (2) can be divided into two parts, the "true" IQ's, which will be written $\underset{\sim}{X}$, and the structural parameters $\underset{\sim}{\beta}$, $\underset{\sim}{\sigma}$, and $\underset{\sim}{\tau}$, and it will be denoted by

$$\underset{\sim}{\theta} = (\underset{\sim}{\beta}, \underset{\sim}{\sigma}, \underset{\sim}{\tau}).$$

Suppose that our prior on $\underset{\sim}{\theta}$ is $f(\underset{\sim}{\theta})$. Some has been said about this prior. However, for the argument below, f will be left unspecified. The joint density of all the observations and parameters is

$$(3)\quad f(\underset{\sim}{X},\underset{\sim}{\theta},\underset{\sim}{Z}) = f(\underset{\sim}{Z}|\underset{\sim}{X},\underset{\sim}{\theta})\ f(\underset{\sim}{X}|\underset{\sim}{\theta})f(\underset{\sim}{\theta}).$$

Therefore, using Bayes theorem, the posterior distribution of the parameters given the data is

$$f(\underset{\sim}{X},\underset{\sim}{\theta}|\underset{\sim}{Z}) = \frac{f(\underset{\sim}{Z}|\underset{\sim}{X},\underset{\sim}{\theta})f(\underset{\sim}{X}|\underset{\sim}{\theta})f(\underset{\sim}{\theta})}{\int f(\underset{\sim}{Z}|\underset{\sim}{X},\underset{\sim}{\theta})f(\underset{\sim}{X}|\underset{\sim}{\theta})f(\underset{\sim}{\theta})\,d\underset{\sim}{\theta}\,d\underset{\sim}{X}}$$

$$\propto f(\underset{\sim}{Z}|\underset{\sim}{X},\underset{\sim}{\theta})f(\underset{\sim}{X}|\underset{\sim}{\theta})f(\underset{\sim}{\theta}).$$

In one sense, the posterior distribution (4) gives our new opinion, after taking the data into account, about all the phenomena under study. However, for our data this distribution is almost impossibly multidimensional, since $\underset{\sim}{X}$ has almost 7000 components and $\underset{\sim}{\theta}$ has over 50 components. Therefore, we chose to consider the marginal posterior distribution of $\underset{\sim}{\theta}$ given $\underset{\sim}{Z}$:

$$f(\underset{\sim}{\theta}|\underset{\sim}{Z}) = \int f(\underset{\sim}{X},\underset{\sim}{\theta}|\underset{\sim}{Z})\,d\underset{\sim}{X}$$

$$\propto \int f(\underset{\sim}{Z}|\underset{\sim}{X},\underset{\sim}{\theta})f(\underset{\sim}{X}|\underset{\sim}{\theta})f(\underset{\sim}{\theta})\,d\underset{\sim}{X}$$

$$(5) = f(\underset{\sim}{\theta}) \int f(\underset{\sim}{Z}|\underset{\sim}{X},\underset{\sim}{\theta})f(\underset{\sim}{X}|\underset{\sim}{\theta})\,d\underset{\sim}{X}$$

$$= f(\underset{\sim}{\theta})\ f(\underset{\sim}{Z}|\underset{\sim}{\theta}).$$

Since $f(\underset{\sim}{Z}|\underset{\sim}{X},\underset{\sim}{\theta})$ is assumed to be normal and linear in the mean in $\underset{\sim}{X}$, and since $f(\underset{\sim}{X}|\underset{\sim}{\theta})$ is again normal, the above integral is normal and can be computed by inspection as follows.

Let

$$\epsilon_i^1 = X_i^1 - \underset{\sim}{W}_i^1\,\underset{\sim}{\beta}^1,$$

$$\epsilon_i^j = X_i^j - X_i^{j-1} - \underset{\sim}{W}_i^j\,\underset{\sim}{\beta}^j, \qquad j=2,3,4.$$

Then the ϵ's are normal and independent with zero mean and variances $(\sigma^1, \sigma^2, \sigma^3, \sigma^4)$. Substituting into (1) and transforming,

$$Z_i^1 = \underset{\sim}{W}_i^1\,\underset{\sim}{\beta}^1 + u_i^1 + \epsilon_i^1,$$

$$(6)\quad \Delta Z_i^j = Z_i^j - Z_i^{j-1} = \underset{\sim}{W}_i^j\underset{\sim}{\beta}^j + u_i^j - u_i^{j-1} + \epsilon_i^j\ ,$$

$$j=2,3,4;$$

i.e., changes in observed IQ scores are functions of environmental factors.

Let

$$\underset{\sim}{Z}_i^{\Delta} = (Z_i^1,\ Z_i^2 - Z_i^1,\ Z_i^3 - Z_i^2,\ Z_i^4 - Z_i^3)$$

and

$$\underset{\sim}{m}^i = (\underset{\sim}{W}_i^1\underset{\sim}{\beta}^1,\ \underset{\sim}{W}_i^2\underset{\sim}{\beta}^2,\ \underset{\sim}{W}_i^3\underset{\sim}{\beta}^3,\ \underset{\sim}{W}_i^4\underset{\sim}{\beta}^4).$$

Then

$$(7)\qquad \underset{\sim}{Z}_i^{\Delta}|\underset{\sim}{\theta} \sim \mathcal{N}(\underset{\sim}{m}^i,\ \underset{\sim}{V}),$$

where

$$\underset{\sim}{V} = \begin{bmatrix} \tau^1+\sigma^1 & -\tau^1 & 0 & 0 \\ -\tau^1 & \tau^2+\tau^1+\sigma^2 & -\tau^2 & 0 \\ 0 & -\tau^2 & \tau^3+\tau^2+\sigma^3 & -\tau^3 \\ 0 & 0 & -\tau^3 & \tau^4+\tau^3+\sigma^4 \end{bmatrix}.$$

Notice that (7) performs the integration in (5) painlessly as the convolution of two normal distributions.

The combination of measurement error and process error in our model makes it a special case of models involving unobservable variables (see Goldberger (1973), Griliches (1973), and Joreskog (1970), and the references cited there). One distinction between their approach and ours is that we can examine the posterior distribution of the unobserved

variables. One way of doing that in this case is to calculate

$$f(\underset{\sim}{X}|\underset{\sim}{Z}) = \int f(\underset{\sim}{X},\underset{\sim}{\theta}|\underset{\sim}{Z})d\underset{\sim}{\theta}$$

and to note that the posterior on the students' intelligences $\underset{\sim}{X}$ will be approximately independent over students. This possibility, although interesting, is not pursued further here.

3. ESTIMATION

Estimation of the parameter space $\underset{\sim}{\theta} = (\underset{\sim}{\beta},\underset{\sim}{\sigma},\underset{\sim}{\tau})$ is based on the fact that the system of equations (6) (or (7)) is in the form of four seemingly unrelated regression equations. Were the covariance matrix $\underset{\sim}{V}$ completely general, it would be exactly in the form studied by Zellner (1962). However the zeros in the upper-right and lower-left corners of $\underset{\sim}{V}$ pose a problem not explicitly considered there.

Zellner proposes that each equation be estimated separately using ordinary least squares, yielding consistent but asymptotically inefficient estimates of the β's. The residuals from these regressions can then be used to obtain consistent estimates of the covariance matrix. Finally, use of the estimate of the covariance matrix thus obtained in a generalized least-squares framework yields consistent and asymptotically efficient estimates of the β's.

Use of this method on the system (6) will also yield consistent and asymptotically efficient estimates of the β's because the estimate of $\underset{\sim}{V}$ will be consistent under the model (6). Alternatively, using the first round residuals to estimate the diagonal elements and elements just off the diagonal of $\underset{\sim}{V}$, and zero to estimate the other elements of $\underset{\sim}{V}$, is also consistent; hence the resultant β's from the application of generalized least squares also are consistent and asymptotically efficient. This second alternative seems to us more in keeping with the model, so we estimated it that way.

All the parameters of the system (6) are identified except for σ^4 and τ^4. However the sum $\sigma^4+\tau^4$ is identified. (See Kadane(1972) for an explanation of identified functions on the parameter space.)

4. IMPLEMENTATION OF THE MODEL

The data were collected in the summer of 1971 from the cumulative school records of all students who had just finished the ninth grade in the Pittsburgh public school system. The time period examined is nine years between 1962 and 1970, during which time a proportion of the group passed from kindergarten to eighth grade in the Pittsburgh system. IQ tests were administered during this period to children in kindergarten, fourth, sixth, and eighth grades.

The tests administered were the Detroit (kindergarten), Kuhlmann-Anderson (fourth grade), Otis Beta FM (sixth grade) and Otis Lennon (eighth grade).

3762 children took at least one IQ test, and 2,067 children took all four tests. This latter group excludes children assigned to special education classes for the slow learner, since such children are not given these IQ tests after they are assigned to such classes. It also excludes children who moved into or out of the school system. These students may have been exposed to different cultural influences than those who were enrolled in the school system for the full nine years. The applicability of our results to children who have moved and slow learners is a topic for future research. We used only the records of the 1713 children which are complete on all the independent variables.

Table 1 (see top of next page) lists the variables used in W_1^1, with their means and standard deviations.

The Sex variable is scored 2 for female, 1 for male. SES is measured by the Hollingshead (1957) Two Factor Index of Social Position, which assigns each individual an index value according to occupation and education (with occupation weighted more heavily). Hollingshead (1957:10) suggests that social class position be determined on the basis of index score as follows.

Table 2

Relation of Social Class to SES Index as Suggested by Hollingshead

Social Class		Range of SES Scores
Upper	I	11-17
	II	18-27
	III	28-43
	IV	44-60
Lower	V	61-77

Notice that the higher the Social Class, the lower the SES index. SES of peers is the average SES for all kindergarteners in the school of the child. Because Pittsburgh in 1962 had neighborhood kindergartens, we take this variable to represent the SES of the neighborhood the child was raised in. The head of household variable is scored -1 if both parents are in the house, +1 otherwise. Race of student is scored 0 for white and 1 for non-white. Non-whites in Pittsburgh are almost entirely

Table 1. Variables Used in Kindergarten Equation

	Variable Name	Mean	Standard Deviation
1.	Constant	1.0	0.0
2.	Sex	1.51	0.50
3.	Number of Siblings	2.74	2.06
4.	SES of parents	53.02	14.32
5.	SES of peers	53.24	9.23
6.	Head of Household	-.75	.65
7.	Race of Student	.380	.486
8.	% Black in School	35.4	39.0
9.	Race · % Black	29.3	40.8
10.	(% Black in School)^2	2775.	3786.
11.	Race · (% Black)^2	2521.	3823.

black. % black is the school average of the race variable, multiplied by 100. Thus the proportion of non-whites in our sample (38%) approximates the sample average proportion of non-whites in the school (35.4%). If each school had the same proportion of blacks, the standard deviation of percent black in school would be zero. In a completely segregated system which has a school average of 35.4% blacks, the standard deviation would be $\sqrt{(35.4)\cdot(64.6)} = 47.8$. Thus the actual standard deviation of 39.0 is evidence of a high degree of segregation.

Table 3 gives a cross-tabulation of SES with race for the entire group of 2,067 students.

Table 3

Cross-Tabulation of
Students by SES and Race

Index Score of Social Position

Race	11-37	38-57	58-77	TOTAL
Blacks Number	30	182	663	875
% of Blacks	3	21	76	100
Whites Number	268	502	422	1192
% of Whites	23	42	35	100

Table 3 shows that there is a relationship between race and SES, with blacks having higher SES, and hence lower class, than whites. In the group of 1713 children chosen for intensive analysis, the correlation between race and SES of parents is .41.

The remaining three variables are higher order terms and interactions of the previous ones.

Table 4 lists the variables used in W_i^2, with their means and standard deviations (see top of next page for Table 4).

Variables 1, 2, 3, 4, 6, and 7 are the same as in kindergarten. However because school-mates need not be the same as in kindergarten, variables 5, 8, 9, 10 and 11 are not the same. Variable 3 is actually the number of siblings when the child entered kindergarten, and for that reason does not change in 4th grade. Variable 12 is the student-faculty ratio of the school of the child, averaged over the five years from kindergarten to 4th grade. Variable 14 is the change in the percent of blacks in the school from kindergarten to fourth grade, and variable 15 is variable 14 times variable 7.

The variables used in the sixth and eighth grade equations were the same as in the fourth grades, and are given below in Table 5.

Table 4. Variables Used in 4th Grade IQ Equation

	Variable Name	Mean	Standard Deviation
1.	Constant	1.0	0.0
2.	Sex	1.51	0.50
3.	Number of Siblings	2.74	2.06
4.	SES of parents	53.02	14.32
5.	SES of peers	53.33	9.37
6.	Head of Household Missing	-.75	.65
7.	Race of Student	.380	.486
8.	% Black in School	35.5	38.4
9.	Race · % Black	29.5	40.5
10.	(% Black)^2	2735.	3711.
11.	Race · (% Black)^2	2514.	3759.
12.	Student-faculty ratio, K to 4	32.5	2.96
13.	# changes of school, K to 4	1.89	1.03
14.	Δ % Black, K to 4	.26	18.2
15.	Race · (Δ% Black, K to 4)	.21	14.4

Table 5. Variables used in 6th and 8th Grade Equations

	Variable Name	6th Mean	6th Standard Deviation	8th Mean	8th Standard Deviation
1.	Constant	1.0	0.0	1.0	0.0
2.	Sex	1.51	.50	1.51	.50
3.	Number of Siblings	2.74	2.06	2.74	2.06
4.	SES of parents	53.02	14.32	53.02	14.32
5.	SES of peers	53.21	9.30	53.30	9.09
6.	Head of Household Missing	-.75	.65	-.75	.65
7.	Race of Student	.380	.486	.380	.486
8.	% Black in School	35.9	39.8	34.9	35.7
9.	Race · % Black	29.8	41.6	26.7	38.9
10.	(% Black)^2	2878.	3913.	2490.	3617.
11.	Race · (% Black)^2	2619.	3942.	2220.	3688.
12.	Student-faculty ratio	29.1	3.69	23.4	4.41
13.	# changes of school	.13	.37	.81	.58
14.	Δ% Black	.45	13.4	-.69	17.64
15.	Race · Δ% Black	-.49	11.2	-1.97	13.87

5. RESULTS

$\hat{V}$, the estimate of the covariance matrix V, is a consistent estimate of V under this model. We obtained

$$\hat{V} = \begin{bmatrix} 236.2 & -145.7 & 0 & 0 \\ -145.7 & 168.4 & -27.3 & 0 \\ 0 & -27.3 & 65.6 & -21.8 \\ 0 & 0 & -21.8 & 61.2 \end{bmatrix}.$$

From $\hat{V}$, the following consistent estimates can be derived:

$$\hat{\tau}^1 = 145.7 \qquad \hat{\sigma}^1 = 70.5$$

$$\hat{\tau}^2 = 27.3 \qquad \hat{\sigma}^2 = 0$$

$$\hat{\tau}^3 = 21.8 \qquad \hat{\sigma}^3 = 16.5$$

$$\sigma^4 + \tau^4 = 39.4$$

The first thing that strikes one about these estimates is that $\hat{\sigma}^2$ is surely too low, that $\hat{\sigma}^1$ is probably too low, and that both are consequences of $\hat{\tau}^1$ being too high. Were $\hat{\tau}^1$ close to the anticipated value of 25 or so, $\hat{\sigma}^1$ would be close to 200, and $\hat{\sigma}^2$ would be about 120, which is high but not unreasonable. These results reveal a weakness in our model. Quite possibly there is non-independence between ε^1 and ε^2, u^1 and u^2, or between the ε's and u's. We leave these possibilities as topics for future research. Any fuller parametrization of V involving zeros where we have put them will lead to $\hat{V}$ being a consistent estimate for V, and hence our estimates of the regression coefficients would still be consistent and asymptotically efficient. As a result, despite this weakness in the model we think the regression coefficients given in Table 6 may be of some interest.

Caution should be exercised in the interpretation of the race and percent-blacks-in-school variables because of the presence of higher-order terms in different ways below.

A few things stand out from Table 6. First, the results on the sex variable indicate that women have an advantage through 4th grade which is lost by the time 8th grade is completed. This is in accord with literature that women mature physically more rapidly than men, although the faster pace of loss between 6th and 8th, compared to 4th to 6th, indicates the possibility of negative relative conditioning of women around intellectual matters.

To help the reader understand which coefficients are important and which are not, we calculate below in Tables 7 to 10 the predicted IQ of a white student

Table 6. Regression Estimates and Estimated Deviations

	Variable	Equation 1 Est.	Equation 1 s.d.	Equation 2 Est.	Equation 2 s.d	Equation 3 Est.	Equation 3 s.d.	Equation 4 Est.	Equation 4 s.d.
1.	Constant	128.4	2.84	-10.24	3.40	10.96	2.28	2.02	2.18
2.	Sex	2.44	.745	1.28	.630	-1.621	.394	-1.80	.380
3.	# siblings	-.474	.188	-.168	.159	-.126	.0996	-.0302	.0964
4.	SES parents	-.208	.033	.0696	.0280	-.0282	.0178	-.0439	.0168
5.	SES peers	-.148	.061	-.124	.0479	-.0832	.0324	-.0192	.0318
6.	Head of household missing	-.747	.578	.229	.490	.256	.306	-.181	.295
7.	Race	-7.04	3.02	9.68	2.77	-1.84	1.58	.430	1.53
8.	% Black	-.00078	.0700	.0987	.0588	.0112	.0389	-.0623	.0391
9.	Race · % Black	-.00391	.118	-.264	.103	.0363	.0668	-.0807	.0653
10.	(% Black)^2	-.00009	.00090	-.00108	.00080	-.00005	.00050	.00041	.00059
11.	Race · (% Black)^2	-.00003	.00116	.00239	.00102	-.00037	.00067	.00074	.00072
12.	St/fac ratio			.0970	.0752	-.00895	.0559	.0481	.0486
13.	# school changes			.0593	.207	-.540	.467	.109	.334
14.	Δ(% Black)			.0259	.0206	-.0235	.0256	-.0234	.0177
15.	Race · [Δ(% Black)]			-.0158	.0260	.0350	.0298	.0439	.0229

with all exogenous variables at the mean for a white student and the predicted IQ for a black student at the mean for blacks. An alternative method of analysis would be to compute significance levels for the estimates. While this latter method of analysis is popular, it is also misleading (Kadane (1973)). For this reason we choose to weight most heavily the analysis of Tables 7 to 10.

Table 7. Effect of Kindergarten Regression Coefficients on Mean Black and Mean White Student

	Variable	Regression Coef.	Black Mean	White Mean	Contribution to Black Score	Contribution to White Score	Δ
1.	Constant	128.4	1	1	128.4	128.4	0
2.	Sex	2.44	1.54	1.49	3.76	3.64	-.12
3.	# Siblings	-.474	3.29	2.412	-1.56	-1.14	.42
4.	SES parents	-.208	60.47	48.547	-12.58	-10.10	2.48
5.	SES peers K	-.148	60.25	49.046	-8.92	-7.26	1.66
6.	Head of Household Missing	-.747	-.663	-.813	.50	.61	.11
7.	Race	-7.04	1	0	-7.04	0	
8.	% Black	-.00078	77.17	9.806	-.06	-.01	
9.	Race · % Black	-.00391	77.17	0.0	-.30	0	8.14
10.	(% Black)^2	-.00009	6634.06	402.77	-.59	-.04	
11.	Race · (% Black)^2	-.00003	6634.06	0.0	-.20	0	
		Total			101.41	114.10	12.69

Table 8. Effects of Regression Coefficients of Change from Kindergarten to 4th Grade on Mean Black and Mean White Student

	Variable	Regression Coef.	Black Mean	White Mean	Contribution to Black Score	Contribution to White Score	Δ
1.	Constant	-10.238	1	1	-10.24	-10.24	0
2.	Sex	1.28	1.54	1.49	1.97	1.91	-.06
3.	#Siblings	-.168	3.29	2.412	-.55	-.41	.14
4.	SES parents	-.0696	60.47	48.547	-4.21	-3.38	.83
5.	SES peers 4	-.124	60.30	49.143	-7.48	-6.09	1.39
6.	Head of Household Missing	.229	-.663	-.813	-.15	-.19	-.04
7.	Race	9.68	1	0	9.68	0	
8.	% Black K-4	.0987	77.64	9.608	7.66	.95	
9.	Race · % Black	-.264	77.64	0	-20.50	0	-4.18
10.	(% Black)^2	-.00108	6615.23	351.98	-7.14	.38	
11.	Race (% Black)^2	.00239	6615.23	0	15.81	0	
12.	Student/Fac. Ratio K-4	.0970	31.24	33.299	3.03	3.23	.20
13.	# Changes in Sch. K-4	.0593	2.13	1.733	.13	.10	-.03
14.	Δ% Black K-4	.0259	.545	-.93	.01	.02	.01
15.	Race · (Δ% Black)	-.0158	.545	0.0	-.01	0	.01
		Total			-11.99	-13.72	-1.73

Table 9. Effect of Regression Coefficients of Change from 4th to 6th Grade on Mean Black and Mean White Student

	Variable	Regression Coef.	Black Mean	White Mean	Contribution to Black Score	Contribution to White Score	Δ
1.	Constant	10.96	1	1	10.96	10.96	0
2.	Sex	-1.621	1.54	1.49	-2.50	-2.42	.08
3.	# Siblings	-.126	3.29	2.412	-.41	-.30	.11
4.	SES parents	-.0282	60.47	48.547	-1.71	-1.37	.34
5.	SES peers 6	-.0832	59.82	49.243	-4.98	-4.10	.88
6.	Head of Household Missing	.256	-.663	-.813	-.17	-.21	-.04
7.	Race	-1.84	1	0.0	-1.84	0	
8.	% Black 5-6	.0112	78.49	9.822	.88	.11	
9.	Race · % Black	.0363	78.49	0.0	2.85	0	1.07
10.	(% Black)^2	-.000048	6891.77	412.65	-.33	-.02	
11.	Race · (% Black)^2	-.000369	6891.77	0.0	-2.54	0	
12.	Student/Fac. Ratio 5-6	-.00895	27.26	30.314	-.24	-.27	-.03
13.	# Changes in Sch. 4-6	-.540	.170	.109	-.09	-.06	.03
14.	Δ(% Black) 4-6	-.0235	-1.29	1.538	-.03	-.04	-.01
15.	Race·[Δ(%Black)]	.0350	-1.29	0.0	.05	0.0	-.05
		Total			-0.1	2.28	2.38

Using the Δ column especially, one can see that some variables do not matter much in their contribution to the explanation of differences between black and white IQ scores, while others matter a great deal. We have lumped all of the variables dealing with race and integration together.

We find Table 11 below to be an informative summary of Tables 7 to 10. In it, we calculate cumulative effects rather than the effects due to differences, and we lump the two SES variables together.

Table 11. Cumulative Effects of SES versus Race-Segregation on the Difference in IQ Between a Mean White and a Mean Black

	K	4th	6th	8th
SES	4.14	6.36	7.58	8.30
Race-Segregation	8.14	3.96	5.03	7.29
Net of Others	.41	.64	.53	.88
Total	12.69	10.96	13.14	16.47

Thus the SES variables account for about a third of the difference at kindergarten, and for more than half the difference at 4th grade and beyond. Note that these calculations are done for fictional persons: a black whose demographic and environmental variables are at the mean for all blacks, a white whose demographic and environmental variables are at the mean for all whites.

Finally, we present a highly tentative analysis of the linear and quadratic terms of the degree of integration variables (% Black) from Table 6. Again we use the cumulative effects, which we compute separately for whites and blacks.

Table 10. Effect of Regression Coefficients of Change from 6th to 8th Grade on Mean Black and Mean White Student

	Variable	Regression Coef.	Black Mean	White Mean	Contribution to Black Score	Contribution to White Score	Δ
1.	Constant	2.02	1	1	2.02	2.02	0
2.	Sex	-1.80	1.54	1.49	-2.77	-2.68	.09
3.	# Siblings	-.0302	3.29	2.412	-.10	-.07	.03
4.	SES parents	-.0439	60.47	48.547	-2.65	-2.13	.52
5.	SES peers 8	-.0192	59.84	49.369	-1.15	-.95	.20
6.	Head of Household Missing	-.181	-.663	-.813	.12	.15	.03
7.	Race	.430	1	0	.43	0	
8.	% Black 7-8	-.0623	70.18	13.375	-4.37	-.83	
9.	Race · % Black	-.0807	70.18	0	-5.66	0	2.26
10.	(% Black)^2	.000406	5842.23	448.93	2.37	.18	
11.	Race · (% Black)^2	.000739	5842.23	0	4.32	0	
12.	Student/ Fac. Ratio 7-8	.0481	21.52	24.555	1.04	1.18	.14
13.	# Changes in Sch. 6-8	.109	1.654	.735	.18	.08	-.10
14.	Δ(% Black) 6-8	-.0234	-5.18	2.310	.12	.05	-.07
15.	Race · [Δ(% Black)]	.0439	-5.18	0.0	-.23	0	.23
		Total			-6.33	-3.0	3.33

Table 12. Cumulative Effects of Degree of Segregation/Integration on the IQ's of Black and White Students

	Linear	Quadratic	Best	Worst	Maximum effect
Whites					
Kindergarten	-0.00078	-0.00009	0	100	.98
Fourth grade	.0979	- .00117	41.8	100	3.95
Sixth grade	.1091	- .00122	44.7	100	3.73
Eigth grade	.0468	- .00081	28.9	100	4.10
Blacks					
Kindergarten	- .00469	- .00012	0	100	1.67
Fourth grade	- .170	.00119	0	71.4	6.07
Sixth grade	- .122	.00077	0	79.2	4.83
Eighth grade	- .265	.00192	0	69.0	9.14

The magnitude of the effect at kindergarten is small and can be disregarded. But the effect of segregation grows, becoming very serious indeed for blacks, especially by eighth grade. Because we are aware of the highly contentious area these results have led us to, we emphasize that these calculations are highly tentative and speculative. One reason we are unsure of these results is that in as highly segregated a system as Pittsburgh had, we have little data for blacks in mainly white schools and vice-versa. This led to large standard deviations, especially on the quadratic terms. The optima are the ratio of the linear term to twice the quadratic, and thus the uncertainty is magnified. Perhaps new data gathered on students who have been through a more integrated school experience would help us estimate these effects better.

6. CONCLUSION

There are several kinds of conclusions to this paper. One is the specific interpretation of this data set given in section 5. A second is that the kind of model we have used can be used to ascertain the effect of any environmental change on school children s IQ. For example, some schools have experimented with open classrooms; this kind of analysis would be appropriate for finding out what effect such a change would have.

We intend to explore several kinds of further analyses on this data set. First, we plan to find out what we can do to raise our effective sample from 1713 to 2067 by doing something about missing independent variables. Second, we would like to include an analysis of achievement test scores, and data on tardiness, absence, health and behavior marks, and grades. All these variables should be endogenous, and perhaps should also enter the IQ equations. Third, we would like to look further into the variance-covariance matrix estimation. Fourth, it would be nice to have variables for the sex of teachers, and to estimate teacher quality. Also, we plan to re-estimate the parameters using the maximum likelihood method. Finally, we could investigate the estimates of true IQ s, the X's, induced by our model. Perhaps in a few years' time we might collect a similar body of data again, now that integration is more widespread in Pittsburgh. It would be interesting to see if its effects are predicted well by our model.

References

Goldberger, Arthur S. 1974 "Unobservable Variables in Econometrics," pp. 193-213 in Frontiers of Econometrics, Paul Zarembka, editor, Academic Press.

Griliches, Zvi. 1973 "Errors in Variables and Other Unobservables," Unpublished Discussion Paper.

Hollingshead, August B. 1957 "Two Factor Index of Social Position." Published by the Author, New Haven, Connecticut.

Joreskog, K. G. 1970 "A general method for analysis of covariance structures," Biometrika, 57, 239-251.

Kadane, Joseph B. 1972 "Identification of Functions of Parameters" Unpublished Technical Report
1973 Review of "Significance Tests" and "Statistics with a View toward Applications" Journal of the American Statistical Association 68, 1025.

Sanday, Peggy R. 1972a "On the Causes of IQ Differences Between Groups with Implications for Social Policy," Human Organization, Winter, Vol. 31: 411-424.
1972b "An Alternative Interpretation of the Relationship Between Heredity, Race, Environment, and IQ," Phi Delta Kappan, December, 250-254.
1972c "A Model for the Analysis of Variation in Measured Intelligence Between Groups," Mental Tests and Cultural Adaptation, Lee Cronbach and Peter Drenth, Eds, The Hague: Mouton Press.
1973. A Diffusion Model for the Study of the Cultural Determinants of Differential Intelligence. Final Report submitted to the U.S. Department of Health, Education, and Welfare, National Institute of Education.

Zellner, Arnold. 1962 "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Association, 57, 348-368.

* We wish to express our appreciation to Russ Winer, John Snyder, and especially Dan Rosen for computational assistance. Kadane's participation was supported in part by National Science Foundation Grant GS-38609. Sanday's and Staelin's participation was supported in part by the National Science Foundation and the Office of Education.

James Tobin, Yale University

When I first became interested in the negative income tax (NIT) in 1964, I had some hope of seeing it adopted but not very much. The Johnson Administration was divided, but generally unfriendly. H.E.W. was committed to gradual improvement of social insurance and existing programs of categorical assistance. The "war on poverty" was supposedly attacking the educational, economic, and social causes of poverty. The Council of Economic Advisers and Budget Bureau could not have found the money for a negative income tax even if they had been thoroughly convinced of its merits.

Nevertheless there were signs that it was an idea whose time was coming. As the press, the public, the Congress worried more and more about welfare reform, the NIT inevitably came to their attention. Although the NIT is naturally an economists' idea, it began to appeal to some professional social workers disillusioned with categorical public assistance. Two or three Congressmen actually introduced NIT bills.

President Johnson postponed decision, and presumably stilled the disagreements of his advisers, by the customary device of appointing a Commission. Chaired by Ben Heineman, the President's Commission on Income Maintenance Programs diligently studied poverty and public assistance in the United States and came out for a negative income tax. The report is excellent in all respects, but President Johnson was not on hand to receive it and his successor was not greatly interested in the findings of a lame duck commission.

President Nixon was getting advice elsewhere, notably from Pat Moynihan, his first counselor on domestic affairs. In the debate during the previous four years, Moynihan had advocated universal children's allowances and had not been deterred when I and others pointed out how costly and wasteful it was to give money indiscriminately to rich children and poor. Now in the White House, face to face with budgetary realities, he designed and sold the Family Assistance Plan, a reform of the welfare system on some of the principles of NIT.

There were many objectionable features of FAP in its several incarnations: income guarantees inadequate, marginal tax rates too high, childless couples and single adults excluded, rules and administration not integrated with income tax, excessive power left to states. The work-ethic rhetoric which the Administration used as a smokescreen to conceal the fact that it was advocating guaranteed income was disingenuous and often disgusting.

Nonetheless I would have voted for FAP as a step forward, hoping it would not be the last step. I don't know whose fault it is that FAP never got through the Senate. Probably there is blame enough for everyone, both the liberals whom Moynihan scolds and the conservatives whom the White House often appeased but never delivered. It is quite evident that with Moynihan in Cambridge or New Delhi and the lessons of the 1972 campaign learned the Administration was only too happy to drop the whole matter.

Now in 1973 the negative income tax no longer seems like an idea whose time is coming. Maybe its time is past, its tide in the affairs of men ebbed. In the United States, that is. Meanwhile the Conservative government in the U. K. is about to implement a system of cashable tax credits, against the opposition of the Labour party. Here the Presidential campaign of 1972 was, of course, a dreadful setback.

What lessons can we learn from the dismal legislative and political history of tax and welfare reform in recent years?

First, Presidential candidates, especially those challenging an incumbent, cannot write tax legislation during campaigns and should not try. They should confine themselves to critique of the status quo and to general principles of reform. Specific proposals are terribly vulnerable, and the arithmetic of taxes and redistribution is hopelessly confused in campaign rhetoric. Senator McGovern's famous thousand dollar demogrant was originally advanced in the spring of 1973 simply as one of a number of interesting possibilities. Little attention was paid to it until Senator Humphrey made it an issue in the California primary. McGovern then put himself on the defensive by embracing the idea and the specific number much more tightly than he ever had before. Unfortunately his defenses were thin. His staff had developed his ideas on tax and welfare reform with minimal technical assistance, and they improvised confusing and erroneous answers to the many specific and arithmetic questions which arose in the California campaign.

Only afterwards was serious work undertaken to design proposals to carry out the candidate's intent and to demonstrate that his basic proposal was financially feasible -- though not of course just by closing upper-income tax loopholes, as he and his staff sometimes seemed to be saying. The serious designs were too late to undo the political damage, which may have been compounded by the candidate's eventual inglorious withdrawal from the whole issue. In the process, lasting damage was done to the cause which was so inexpertly championed. It will take time and patient persuasive effort to convince people that income guarantees, demogrants, cashable tax credits, negative income taxes, and all that are not crackpot ideas.

Second, I fear one must conclude that the probabilities are against enacting in one magnificent stroke a comprehensive package of tax and welfare reform. The rhythm of American politics seems to provide legislative majorities

for sweeping change and redistribution no more often than once a generation. Consider the periods of drought between the first Wilson administration and the New Deal, and between the New Deal and Johnson's Great Society Congress, whose promise was tragically ended by the escalation of the war in Vietnam.

Proponents of tax reform, discouraged by reversals suffered in the horse-trading negotiations of piecemeal reform, often dream of starting over again from zero. They observe that less than half of national personal income is federally taxable, one obvious reason why tax rates are high. Let everyone toss in his privileges, exclusions, exemptions, deductions, and take his chances on a simple tax on a comprehensive base, with cashable credits for all adults and children. In theory there is a latent majority coalition for a new social financial contract of this kind; winners would be much more numerous than losers. But in practice that coalition has yet to be mobilized. It is too easily splintered by internal conflicts of interest: families versus single individuals, small families versus large; renters versus homeowners; young versus old; poor versus near-poor, and so on.

The normal rule of tax reform is that almost nobody's taxes can be increased. I say "almost" because some loopholes and privileges are so notorious that they are fair political targets. But the list is pretty short, and the revenue involved pretty small. Any major redistribution through the tax system requires cutting into some widespread tax concessions, not generally perceived as outrageous or even unfair. Examples are the favorable treatment of capital gains, philanthropic contributions, and home ownership. Even if these and other erosions of the tax base could be repaired, a major redistributional tax reform requires higher tax rates, and greater liabilities for many taxpayers. Citizens who might accept higher tax liabilities for war or some other substantive national purpose will resent them deeply when they are being openly redistributed to other citizens.

This is why Senator Humphrey's secretary was so damaging to Senator McGovern's demogrant proposal. In a nationally televised California primary debate, Humphrey pointed out that a single secretary earning $8,000 a year would pay $567 more in taxes under McGovern's proposal; the higher tax rate would more than offset her $1000 demogrant.

It was not clear how Humphrey had made this calculation, since no specific McGovern proposal had been set forth. But, although the example may have been exaggerated, it was qualitatively correct. The demogrant proposal did involve a horizontal redistribution from single individuals and couples to large families, along with a vertical redistribution from rich to poor. Never mind that the illustrative secretary was rich as single individuals go -- in the upper 17% of such persons in 1970. Never mind that she personally would, thanks to salary increases, be better off than in 1970 in after-tax income in 1974 or 1975, whether or not the McGovern reform was adopted -- though of course better better off if it was not. The normal growth of after-tax income, with constant tax rates and rules, is not regarded as fair game for additional taxes. The public image was that an ordinary working girl with an income in four figures would be unfairly burdened.

Under these political restrictions, the best that a redistributionist can hope for is to claim some share of the annual fiscal dividend -- the growth in revenues from existing taxes. This is not easy because of the intense budget competition for those funds. With the fiscal dividend, it is possible to decrease the taxes of the poor and to increase their negative taxes, without explicitly damaging Senator Humphrey's secretary or any other taxpayer. The damage to them in tax reductions foregone is a much smaller political obstacle.

The Moynihan-Nixon Family Assistance Plan is an example of incrementalist strategy. However, it was not a strategy which would lead gradually to a more fundamental reform. Even when we are confined to small steps, we should be following a path that leads somewhere. In particular, I think it is desirable to begin making reforms within the framework of the federal income tax, so that we are not forever stuck with a dual system, welfare for the poor, the income tax code for the rest of us.

In this spirit, I would suggest beginning to convert exemptions and deductions into tax credits, cashable to the extent that they exceed tax liabilities. One step, for example, would be to convert personal exemptions of $750 into cashable credits of $375; since almost no one is subject to a marginal tax rate greater than 50%, almost no one would lose. The credits for adults could then be gradually increased. In similar vein, the standard deduction and homeowners' deductions could gradually be transformed into cashable credits. Cashable credits would gradually take the place of public assistance, and in time an integrated system would evolve. Meanwhile, the working poor and near-poor, who are short-changed by our present welfare and tax systems, would be getting the better breaks they so greatly deserve.

Third, a solution must be found for the pyramiding of actual and implicit income tax rates. Benefits under a host of federal and state programs are scaled to income: public assistance, medical care, rent subsidies, food stamps, educational grants, and more. To the marginal income tax rates implicit in these programs may be added regular income taxes and the ever-increasing social security tax on earnings. As a result it is easy to display horror cases where the earning of an extra dollar of income costs a family more than a dollar in benefits lost or taxes due. These cases, or less dra-

matic examples damaging to work incentives, would be more frequent under any welfare reform -- whether F.A.P. or N.I.T. or demogrants-- which would increase the number of families eligible for income-tested cash assistance along with various in-kind benefits. The Senate Finance Committee's ostentatious discovery of this fact was one of the nails in the coffin of F.A.P. It seemed a miscarriage of justice to place the blame on the cash assistance proposal rather than on the proliferation of uncoordinated in-kind programs. Be that as it may, the problem must be faced more squarely than in the past.

The sweeping solution is to supersede in-kind programs with the cash program. In-kind programs like rent subsidies might continue, but the value of the housing benefits would be subtracted from cash benefits due, even if the net result was that the family owed tax. A less drastic solution would charge less than 100% of in-kind benefits against entitlement under the cash program. If 80% of an in-kind program were charged, that program would add only 20% of its implicit tax rate to the overall marginal tax. Escalation of disincentive rates can also be mitigated by treating various assistance programs sequentially, including in the income that determines entitlement to the 3rd kind of assistance all the net benefits received from assistance programs numbers 1 and 2.

Fourth, no new system of federal income guarantees can be expected to finance the benefits which some recipients of public assistance receive in the most generous states and localities. Let the best not be the enemy of the good. It is just not economically or fiscally feasible for New York or Connecticut AFDC benefit levels to be universalized across the whole nation to all categories of families. Sometimes a negative income tax is dismissed on this account -- if it can't even provide income guarantees equal to the best current welfare benefits, what good is it? The answer, of course, is that the income guarantees would benefit millions throughout the country who are not eligible for those higher welfare benefits.

It can be argued that there is in equity an obligation not to reduce the benefits of existing welfare clients. Recognizing this obligation, the federal government should meet the costs. But equity in this sense dictates a grandfather or grandmother clause for individuals, not for categories of individuals or for states and cities. There is no federal obligation to perpetuate existing geographical inequities in welfare benefits, which are in any case an incentive for uneconomic migration and location. Of course any state or city can in its own discretion finance its own cash assistance program or negative income tax.

Fifth, the public's fears that their hard-earned tax dollars may support malingerers and loafers must be allayed if any national system of income guarantees is to be acceptable. It is not enough to build work incentives into the system, in the form of income "disregards" and tolerable marginal tax rates. It is not enough to cite the New Jersey experiment and the other voluminous evidence that there are precious few people who enjoy living idly on handouts. It is not enough to point out that the hard pressed middle income taxpayer should direct his outrage to the idle rich who pay less taxes than they should rather than to the idle poor. Public opinion just won't accept a system under which ablebodied adults may loaf at government expense, and there are bound to be a few examples of some who do.

Various devices -- e.g., registration for work at a local public employment office -- have been suggested and debated. I believe a suggestion by Harold Watts has merit. Let part of the income guarantee (or tax credit) available on account of an adult of working age be contingent on a declaration, under the usual penalties for false statements on tax returns, that he or she was engaged in one or more of the following activities: gainful employment or self-employment, job-seeking, child care and housekeeping, schooling, unpaid volunteer public service. This requirement would not discriminate against the poor; everyone who claims this tax credit, whether he takes it in cash or in reduction of his tax liability, would have to meet the requirements. Nor would a whole family be penalized for the delinquency of one of its adult members; the benefits or tax credits due to the other adults and children would continue.

Sixth, the general public also resents supporting the children of fathers who have deserted them. Men and women who fulfill their own responsibilities as parents don't wish to be burdened with expenses left behind by parents who have abandoned these responsibilities. Worries on this point have some foundation, as indicated by the continuing growth in the number of dependent female-headed households. Current welfare programs contain provisions for seeking out absent fathers and requiring them to contribute to the support of their deserted children. But these provisions have never been very effective. It is fair to say that they have not been popular with social workers, who have seen them as an authoritarian and punitive attempt to impose bourgeois values on the poor. There is justice in the suspicion, but it is unfortunately no trivial matter if the society ends up supporting millions of deserted mothers and children whose fathers are earning comfortable incomes elsewhere.

A possible answer is to assess an extra tax on the income of an absent father or mother for every child he or she is not supporting -- unless of course the obligation has been undertaken by a step-parent or foster parent. To enforce this penalty it would be necessary to assign children social security numbers at birth and to associate them with the numbers of their parents. These social security numbers would also be the basis for claiming NIT benefits, tax credits, or dependents' exemptions on account of children; they would prevent the same child from being claimed as dependent in more than one family.

THE ROLE OF INCOME-CONDITIONING IN THE AMERICAN SYSTEM OF TRANSFERS

Robert J. Lampman, University of Wisconsin

Between 1960 and 1973 social welfare expenditures under federal, state and local government programs increased from $52 billion to $215 billion. They were equal to 11.8 percent of gross national product (GNP) in 1960 and 17.6 percent in 1973. The three leading classifications of these expenditures are social insurance ($86 billion), education ($65 billion), and public aid ($28 billion). Lesser amounts are listed under health and medical programs, veterans programs, "other social welfare," and housing. In addition to these public expenditures, there is a growing set of private transfers to persons in the form of such things as pensions, health insurance benefits, scholarships and charitable grants. This battery of public and private transfers, in cash and in-kind, is financed by taxes, which, in some cases, are designed to further certain transfer purposes, and by private contributions. The public and private components of this American system of transfers now take in and pay out an amount equal to almost one-fourth of GNP. And that share seems destined to grow.

The scope and scale of this system have grown particularly rapidly since 1964, starting with the introduction of medicare and medicaid and federal aid to elementary and secondary education, and continuing in more recent years with other innovations as well as with the expansion of existing programs. I will, in this paper, point to a few of the more remarkable changes of recent years. The changes selected for discussion were all designed to concentrate their benefits on families in relatively low-income status. They all have to do with the income-conditioning of benefits, a practice which is now surprisingly popular.

Three Recent Changes

Supplemental Security Income (SSI), which goes into effect January 1, 1974, is our first nation-wide negative income tax. Perhaps it is the second such plan in the world, following the British Family Income Supplement of 1971. It covers only the aged, blind, and disabled, but it does establish near-poverty-line guarantees ($2500 for a couple) in all states and it sets uniform rules for determining eligibility and benefits. It will be administered by the Social Security Administration and financed out of general revenues. States must contribute enough to maintain present guarantees for current beneficiaries and are encouraged to add to SSI levels for new recipients. The guarantee is reduced dollar for dollar by all but the first $20 per month of non-earned income (including social security benefits) and by 50 percent of earnings after the first $65. In other words, after certain set-asides, the implicit tax rates are 100 percent on non-earned income and 50 percent on earnings. This will produce break-even points in the neighborhood of $6000 for a year for those couples with earnings. There is no work test and no relative responsibility test, but there is a resources test.

SSI will add about $3 billion of cash income to low-income families and individuals in its target categories. This will fill a substantial part of the poverty-income gap which was about $12 billion in 1972 and is probably less in 1973. However, this effect is somewhat muted by the withdrawal of eligibility for foodstamps on the part of most of those claiming SSI.

A second notable change is the liberalization of Aid to Families with Dependent Children (AFDC). The 1967 Amendments specified that the tax rate on earnings cannot exceed 67 percent. Researchers find that because of the set-aside of $30 a month indicated by federal law, and because of practice in some states of ignoring earnings that bring total income up to stated standards, and because of deductibility of work expenses (a favorable ruling requires deduction after the 67 percent tax rate is applied to earnings), and because of variability of rent allowances, the tax rate is rarely as high as 67 percent. This means, of course, that break-even incomes are substantially higher than guarantees. AFDC has also been liberalized by court rulings outlawing rateable reductions in benefits to extend a fixed-sum appropriation through a benefit period, state residence requirements, and state rules on the issues of "a man in the house" and the non-adoptive step-father, as well as the earlier practice in some states of counting "expected" but not actually received contributions from relatives. All of these changes, plus a less harsh stance by administrators and, perhaps, a decline of stigma associated with receiving AFDC benefits, have contributed to a close to 100 percent take-up by eligibles, if we are to believe Census reports of numbers and incomes of broken families. This remains true in spite of the Talmadge Amendment of 1972 which mandated a work test for mothers whose youngest child is six years of age or older.

AFDC guarantee levels have varied widely from state to state, with the highest-paying states' guarantees running six times above those in the lowest. However, the advent of the foodstamp program has served to reduce this variation. AFDC cash plus foodstamp bonus values now yield guarantee levels for a family of four of $2316 in Mississippi and $5046 in Hawaii, i.e., a difference of only about two to one. Similarly, the availability of foodstamps for working poor families diminishes the gap between what a low-wage earner can provide for his family and what they would get if he deserted and let them go on AFDC.

The third remarkable change in our system of transfers is the evolution of foodstamps into a major program. The foodstamp schedule which goes into effect on January 1, 1974 has a guarantee of

$1704 for a family of four, a set-aside of $360 as well as deductibility of taxes paid and of certain working expenses, and a tax rate of 30 percent, leading to a notch where $288 worth of benefits are lost, down to a break-even point of $5676. This schedule is to be operational in all areas of the country by next July. This year, about 12.5 million people have benefited from the program, but the higher benefit schedule and the mandating of it nation-wide will make more than 30 million people eligible, in spite of the fact that most SSI recipients are not eligible. Hence it is a second nation-wide negative income tax, but, in this case, one with benefits in kind. It can be argued that some part of the $5 billion to $10 billion of foodstamp bonuses of the expanded program should be counted as reducing the poverty-income-gap.

All the bonus values going to those with money income below poverty lines would be relevant to this consideration if one could affirm that foodstamps are as good as money, which they are when people would spend at least as much on food as they can claim in foodstamps. This is roughly the case at poverty-line incomes (the poverty-line for a family of four is now $4300) and above. The monthly foodstamp allotment for a family of four is $142. If such a family has a money income of $375, their foodstamp bonus is $38, which yields a total income of $413. At that level of income they are likely to spend $142, or one-third of income, on food. On the other hand, at very low levels of cash income this is not likely to be the case. For example, at $100 of cash income the full foodstamp allotment of $142 would cost the family $25. A family in that situation is unlikely to want to devote $142 out of their full income of $217 ($100 in cash plus $117 of foodstamp bonus), that is, two-thirds of their income, to food. Hence, they are likely to bootleg part of their stamps or food and to lose something in the process, or to buy something less than their full allotment of coupons. The fact that these calculations are based on monthly rather than annual income means that foodstamp benefits are worth more to a family with income that varies from month to month than to one with a stable income.

Even with those limitations, foodstamps will serve as a useful supplement to income for many working-poor families. The low guarantee (relative to the average AFDC guarantee) is partly justified on the ground that intact families (unlike single-parent families) have the option of taking income in the form of home-produced child care at the same time that they get income from the market. As we noted, foodstamps also serve well to narrow the interstate variation of AFDC benefits. However, they heighten the disincentive problem for those on AFDC. Since the foodstamp formula takes account of AFDC benefits, the 30 percent tax rate implicit in the foodstamp schedule does not simply add to the 67 percent rate in AFDC, but it does produce a combined rate of about 77 percent on earnings. Even though, as we noted above, the actual rates of tax in AFDC are below the nominal rates, it is unreasonable to expect that large numbers of welfare mothers are going to work voluntarily in this situation, especially if they have any un-reimbursed work expenses. Even deductibility of child care cost will leave some part of that cost as a tax on earnings. Hence, a woman will have to earn a considerable amount before her "disposable income," net of payment for child care, is equal to the guarantee at zero earnings, in which situation she consumers her own home-produced child care.

There are intriguing questions of equity here. Some single-parent families with earnings below AFDC break-even levels (which range up to $8000 and above in some states) are ineligible for AFDC help because their earnings are above the guarantee level. (The H.R. 1 Family Assistance Plan would have corrected this anomaly.) This produces understandable claims by those excluded from AFDC for help in meeting their child care costs, and Congress has responded by liberalizing income tax deductions for child care and by pushing for direct government provision of day care on a sliding scale of benefits, with partial subsidy extending above median family income levels. Currently, federal support of day care runs to about $2 billion a year, with added amounts via welfare deductibility and tax deductibility of day care costs paid by families.

Cumulation of Tax Rates

However, to return to the incentive issue, the fact is that AFDC, foodstamps, and the unreimbursed portion of work expenses, leave many welfare mothers, who are now the majority of women heading families with children, living under a regime of 100 percent implicit tax rates. This is without taking account of medicaid, which varies considerably from state to state, but which has inequitable and disincentive features in it. The Nixon Administration, according to newspaper reports, is going to propose again what they did in 1971, namely, the nationalization of medicaid and a more orderly income-conditioning of its benefits without regard to welfare status. However, a family would lose eligibility if the head is regularly employed fulltime, in which case they would have the lesser protection of compulsory private insurance contracted for by the employer. Apparently, the medicaid guarantee for a family of five would be on the order of $1000 in insurance terms. This would be accompanied by a zero tax rate on incomes up to a certain low level, with co-payments functioning as an implicit tax rate beyond that. So this solution to the "medicaid mess" would still leave medicaid with a share in the cumulative tax rate burden.

The Administration is also considering another variant of a negative income tax in the form of an income-conditioned housing allowance, in which the guarantee would vary by family size and also by locational difference in the cost of decent housing. Moreover, the Administration has pushed for, and the Congress has authorized but not yet funded, a negative income tax in the form of Basic Opportunity Grants (BOG) for college students. The guarantee is equal to one-half the cost, including living costs, of attending a

college, up to $1400. Beyond certain set-asides, the tax rate is 20 percent if the student is a family-dependent for income tax purposes, and 75 percent if he is independent. The break-even income level for a family of four with a dependent student is above $10,000; it is $2300 if the student is independent.

Let us assume that the proposed revision of medicaid, the housing allowance, and the BOG program all come into being and take their place alongside income-conditioned child day care, foodstamps, AFDC, and SSI. Some beneficiaries of some of these benefits will at the same time be paying payroll and income taxes and will be bearing unreimbursed work expenses. As we indicated before, AFDC recipients will typically confront cumulative tax rates of 100 percent or so in certain ranges, even without reference to medicaid, housing allowance or BOG. But what about low-income people who are not on AFDC or SSI? They will face the 30 percent tax rate in foodstamps; a tax rate of, say 20 percent in the housing allowance; possibly, depending upon employment status, a 20 percent rate in medicaid; and, depending on circumstances, some combination of tax rates from among the following: child care benefits, BOG, unreimbursed work expenses, payroll taxes, and income taxes. Hence, it would appear that many working poor and near-poor family heads will face cumulative tax rates well above 50 percent in certain income ranges and at certain stages in the life cycle. Thus, we would seem to be on a collision course with the expectation that most people should have strong monetary incentive to work.

Let me re-state what the cumulative tax rate problem is. If the same earnings are taxed twice by, say, a payroll tax and an income tax, the two tax rates are added together to determine the combined tax rate. But here we are also talking about implicit tax rates, that is, the rates of reduction of a cash or in-kind guarantee. If there are two such guarantees, both subject to a 50 percent tax rate, and if the break-even points are the same, the combined implicit tax rate is 100 percent. As we have seen, the combining of positive tax rates and implicit tax rates, the latter associated with negative income taxes or income-conditioned benefits, has gone some distance. There are only two basic ways to back off from the prospect of high cumulative tax rates. One way is to reduce the combined guarantee and the other is to extend the break-even points of some or all of the benefits. Neither is likely to be happily received by everybody. The first means reducing benefits for poor people; the second means raising taxes.

Ways to Reduce Combined Tax Rates

Ways to limit the combined guarantee include the following. Legislation can require that the benefit from one program be subtracted from the guarantee of another. Thus, social security benefits must be subtracted from the guarantee of SSI. Another rather clever way to limit the combined guarantee is to consolidate two programs by folding the current budget of an as yet undeveloped program into the proposed budget of another and producing a new combined guarantee which is actually less than the formulas would dictate. This seems to be how foodstamps were "cashed out" for SSI recipients. A different way is to count the benefits of one program as "income" in computing the benefits of the next. Thus, AFDC benefits are "countable income" for foodstamps. Another technique is to design benefits to avoid simultaneous receipts. Legislation could specify, for example, that anyone claiming foodstamps is ineligible for a housing allowance, or, legislators could simply anticipate that few families will claim income-conditioned pre-school and college aid at the same time.

The other basic tactic for avoiding high cumulative tax rates is to extend break-even points of one or more of the benefits. Set-asides, disregards, and deductibles will accomplish this. Another way is to simply not income-condition the benefit at all and to let the break-even point be determined by the tax system. The model here is public education. Perhaps the next best candidate for this kind of treatment is medical care benefits. A way to at least confuse the break-evens is to use a different income accounting period for each benefit. Some may use a month, some a quarter, some a year. BOG would use the income of the prior year as the base for the current year's benefit. The British Family Income Supplement uses an estimate of future earnings as the base. Still a different way to extend a break-even is to allow considerable flexibility in the definition of the "family" whose income is to be counted. Congress tried to confine college students applying for foodstamps to the families who claimed them as income tax dependents, only to have the Supreme Court find such a restriction in violation of Constitutional due process. (See U.S. Department of Agriculture v. Murry, 41 LW 5099, U.S. Sup.Ct. No. 72-848, June 23, 1973). It would seem likely that this same reasoning would apply to the income tax dependency test in the BOG scheme. If it were, many college students from affluent families would go "independent" and claim the maximum grant. This would only be an extension of a profound trend in welfare law toward narrowing family responsibility. SSI cuts the few remaining legal ties between children and their aged parents. Note that in the cases cited, guarantees and nominal tax rates are unchanged, but people behave in such a way as to increase the number of beneficiaries and the cost to the treasury. High cumulative tax rates will encourage such behavior.

The set of in-kind income-conditioned benefits now in place and on the horizon seem to leave little room for a cash benefit for the non-categorical or working poor. Even a plan with an implicit tax rate of about one-third--like the McGovern plan or the proposed British tax credit scheme, which, incidentally has no guarantee for those earning less than one-fourth the average wage--would appear to be rivalrous with foodstamps, medicaid, a housing allowance, and other benefits and positive rates we have mentioned. Some have advocated an earnings subsidy as a way out of this problem. Thus, the Senate recently passed a bill

to refund to workers the social security tax (both that paid by the employer and the employee) on earnings up to $4000 (at that point the refund would equal $400) and to diminish the refund to zero at $6000 of earnings. This would offset to a minimal degree the cumulated tax rates listed above with respect to earnings below $4000, but would introduce a new implicit tax rate of 20 percent between $4000 and $6000 of earnings. Another "way out" is a wage rate subsidy, but that is not easy to confine to poor families, it sets up disincentives to taking jobs at higher wages, and it is difficult to administer. Both an earnings subsidy and a wage rate subsidy are antithetical to deductibility of child care expense. The alternatives of subsidizing private employers or public agencies to create jobs for the poor have at least as many problems as earnings and wage rate subsidies.

A Concluding Comment

The recent moves toward more income-conditioning of benefits mean, it seems to me, that advocates of a simple, straight-forward negative income tax with a moderate tax rate in it are caught between a rock and a hard place. They may wriggle out of the difficulty by designing a negative income tax with very low tax rates or by shifting over to an earnings subsidy (which means negative rates). Or, they may try to cancel out or consolidate guarantees or extend the break-evens of some of the non-cash income-conditioned benefits. Stating the alternatives this way is to indicate my belief that we are approaching the outer limits of income-conditioning.

The practice of confining benefits to low-income families is based upon what might be called the doctrine of minimum provision. As we have seen, it seems to have its own internal dynamic. If minimum provision is assured for education, why is it not also for health care, food, housing, pre-school child care, higher education, legal services, and yet other goods and services? The level of minimum provision is often set well above the level that families with median income will voluntarily consume, e.g., federal standards for child care. The only restraints on this dynamic appear to be unwillingness to tax the non-beneficiaries in order to fully fund the high standards for all eligibles, and concern for high cumulative tax rates on beneficiaries.

The next phase in the development of the American system of transfers may see a greater emphasis on two other doctrines that power the growth of the system. These are the doctrine of sharing income loss and the doctrine of sharing in extraordinary expenditures. The recent emphasis on income-conditioning may turn out to be only a chapter in a longer book.

REFERENCES

Henry J. Aaron, Why Is Welfare so Hard to Reform? Washington: Brookings, 1973.

Kenneth Boulding, The Economy of Love and Fear, Belmont, California: Wadsworth, 1973.

U.S. Congress, Joint Economic Committee, Subcommittee on Fiscal Policy, Studies in Public Welfare, Papers numbered 1-12, Washington: Government Printing Office, 1972-1973.

E.R. Fried, et al., Setting National Priorities: The 1974 Budget, Washington: Brookings, 1973.

Howard Glennerster, "A Tax Credit Scheme for Britain--A Review of the British Government's Green Paper," Journal of Human Resources, Fall 1973, Vol. 8, pp. 422-435.

Robert J. Lampman, "Transfer and Redistribution as Social Process," Social Security in International Perspective, Shirley Jenkins, ed., New York: Columbia, 1969.

D. Piachaud, "Poverty and Taxation," The Political Quarterly, January-March, 1971.

J.F. Sleeman, The Welfare State, London: Allen and Unwin, 1972.

Social Security Administration, Research and Statistics Note No. 22, "Social Welfare Expenditures in Fiscal Year 1973," December 24, 1973.

Adrian L. Webb and Jack E.B. Sieve, Income Redistribution and the Welfare State, London: London School of Economics, 1971.

DISCUSSION

ALVIN L. SCHORR, COMMUNITY SERVICE SOCIETY OF NEW YORK

As members of this panel know, I have long thought the negative income tax an ill wind that no one, no matter how dedicated, would blow good. I was, therefore, from a fairly early point, opposed to the President's welfare reform. But I am heartened by Prof. Tobin's and Prof. Lampman's papers which forego arguments about "who did what to whom," to extract lessons from the experience of the last few years. And they are, as usual, thoughtful and practical. So we begin to move forward once more.

I would like to point to one lesson that is, it seems to me, implicit in Prof. Tobin's paper, although at moments he seems to overlook it himself. That is, none of the income maintenance proposals that has recently been put forward is intrinsically efficient or inefficient. Tax credit, negative income tax, children's allowance, welfare reform: All may be efficient or inefficient, depending on design. While he favors a tax credit, Prof. Tobin rejects a children's allowance as inefficient. Yet a children's allowance, if it did away with the tax exemption for children, would probably be more efficient than a tax credit at the same payment level. Conversely, many economists have supported the negative income tax because of its presumed efficiency. Yet the Heineman Commission dutifully reported that a negative income tax with a $3,600 minimum would be only 36 percent efficient.

So one perceives that a scale of incentive payments is one one of various approaches to efficiency. Stigma and repressive administration have been a much favored method in practise, if not in conference papers. Trading off a proposed benefit against an existing tax benefit is a method common to the tax credit and children's allowance. Designing a program for a population group that tends to have a large proportion of poor people is a fourth method and is, as it happens, a principle of social security. (For example, retirement insurance is about 50 percent efficient -- in other words, retirement insurance is more efficient than a poverty-level negative income tax.) I am saying, in short, that if we test efficiency by inspection rather than by authority, we shall find a more versatile set of proposals open to us.

That is fortunate, in the light of the lesson that is Prof. Lampman's carefully developed main point. That is, we have reached and perhaps exceeded the limits of income-conditioning. With tax rates from one program pyramidding on others, incentive to work in any of them may be quite wiped out. Indeed, other problems arise before the problem of incentives. With the proliferation of regulations that relate one program and benefit level to another, they all become confused. That was the fate, in simpler days, of the AFDC work incentive that Prof. Lampman discusses. HEW financed (and suppressed until welfare reform had died anyway) a large-scale study that shows that many recipients did not respond to those work incentives at all. Why not? It seems that neither they nor their income maintenance workers understood the calculation or believed they would really benefit. Nor, some months after researchers had carefully explained the incentives, did they prove more effective. Indeed, I believe that even today investigation would turn up many localities that have never implemented that particular 1967 amendment. They are on an undeclared strike against legislation they fail to understand or regard as hopelessly complex. If even more income-conditioned programs are developed or if we attempt to subordinate the tax rate in one to another, as has been suggested, I suspect that would compound the confusion.

I do have a suggestion regarding this particular problem. It arises out of the observation that pyramidding is not a function of payment arrangements but only of the effort to recapture earned or "excess" income. In other words, it is not paying out, no matter how many programs are involved, that creates the problem, but taxing back or, as the usually understated British say, claw-back. Suppose we relied solely on the graduated income tax for claw-back, and paid out in as many or as few separate programs as we like -- simply taking pains to make all payments taxable? That would be quite workable and a great simplification. The problem, of course, is that public assistance, food stamps, and all the rest have much higher rates of taxation than the income tax at comparable levels. So this complex, probably unworkable system of incentive arrangements and taxation of benefits exists mainly to protect a specially high tax rate for the poor. It is too bad we cannot trifle with that, for otherwise we should have had a solution to this problem.

I have so far offered comments in terms that may interest economists. In this matter of income-conditioning, however, I believe that social issues are far more significant and will have longer term consequences. Our educational system has tended to confine the children of poor people to poverty. In housing and neighborhoods, we separate economic classes more than other industrialized countries. It has lately been argued that we are developing two distinct labor markets. And here, in a transfer system that disposes of almost a fourth of GNP, we also see the deepening of a dual system. It goes without saying that the educational system, residential arrangements, the labor market, and the transfer system interconnect. With all linked and going in the same direction, we may be developing the permanent underclass of which Gunnar Myrdal once warned -- a true duplex society.

A duplex society is not desirable in any country. In the United States, with our ideology of social mobility and with the racial overtones

that class divisions carry, it is explosive. Education, housing, and employment are not our subjects here, and their policies may be more difficult to manage. But transfer policy is directly subject to manipulation. There at least we shall turn away from income-conditioning if we want a nation that is at all at peace with itself, as well as for the reasons that Profs. Tobin and Lampman have offered.

Before I leave income-conditioning, let me speak directly to Prof. Tobin's suggestion of a work declaration. He proposes it, I believe, not so much because it is intrinsically desirable but because he thinks the public thinks it is. The work test he has borrowed from Harold Watts seems so broad as to exclude no one, probably by intention. Who, being otherwise idle, cannot at the least claim to be doing volunteer public service? But the interesting thing is that such a declaration is not required to take a tax exemption under current law, and Prof. Tobin understands better than I that an exemption and a credit are the same money. Then what is different? Why, in our heads we understand that we would be giving these $375 payments to a number of people who are too poor to pay income tax as well as to all the rest.

I will make my point about this in a moment, but should say a word about taxing non-supporting fathers. I have no desire to defend social workers -- we must be almost as guilty as economists of having failed the nation in these desperate years -- but social workers are not the reason fathers don't support. If some institution must be found responsible, it may be the courts and prosecutors. One must say in their defense that they don't enforce support because they find it unreasonable to do so. Most separated and divorced men soon remarry and found new families, and few have incomes adequate to the support of both. (I remember the case of a man who was extradited from Maryland to Connecticut and jailed for non-support. The Connecticut prosecutor made a fine showing, and the second wife in Maryland promptly applied for AFDC -- and received it.) In general, courts with the facts in hand order less in support than welfare departments, for example, tend to require. You may find that hard to believe, but it is so. I am trying to say that the problem about support -- and it is a problem -- lies in deep-rooted American patterns of child-bearing and marriage. It will not be dealt with by nuisance taxes or new administrative devices; and proposing them is not a serious way of treating the problem.

I talk about the work declaration and non-support in the context of income-conditioning because it represents a lesson that has, perhaps, not adequately been learned. That is, once we start to design transfer programs to regulate people's lives, we enter on a slick road to "the welfare mess." The President's proposed welfare reform should be an instructive illustration. It was designed by people who intended the simplest sort of income-tax-type administration. In the hands of Congress and lobbyists intent on dealing with the poor -- or their idea of the poor -- it took on a load of requirements about family relationships, work, training, and child care that could not have been administered at all, let alone simply. I hope you see my point here -- that it is precisely the proliferation of these requirements -- conceived by the mind-set of income-conditioning -- that turns what we call a negative income tax into what we call a welfare mess.

But Prof. Tobin's point is -- certainly Prof. Lampman's point is -- that we should move on from income-conditioned programs. If we really grasp what that change means, we don't need all the talk about work and family breakdown that we have had -- not to justify transfer program proposals, at any rate. One may think the public will require such discussion. I doubt that. I think the people sitting here lead the public, whether for good or error. In any event, we should give the public the benefit of our best thinking, without supposing that they will think the worst.

Such a view seems to me to be highly compatible with what I take to be the most important lesson that Profs. Tobin and Lampman have gleaned. That is that we should keep our eyes on the whole transfer system. We are not, as Tobin says, forever stuck with a dual system. The tax credit he suggests would, at a wild guess, cost $20 billion net. Yet it is only one element of a series of proposals that one would offer. As has been pointed out, they would have a practical advantage. We have learned that there is small chance of wiping out what we have and writing on a clean slate. If we have a versatile arsenal of measures, each of them calculated to favor people at the bottom of the income distribution, we shall have a better chance of succeeding over a period of time. And we shall have much more to succeed about!

I suppose my underlying point, which I think I take from Profs. Tobin and Lampman, is that we are not dealing solely with anti-poverty measures in some simple sense. We are dealing with the distribution of income in the United States, and how it must be altered. It is a difficult, long-term struggle, but that is the struggle.

Discussion

Alice M. Rivlin, The Brookings Institution

Whoever asked three economists and a social worker to play "historian for a day" was a brave man. Inevitably, he has elicited three very different interpretations of the meaning of recent events for the future of income maintenance and welfare reform.

Professor Tobin, in his new role as current historian, reflects on "the dismal legislative political history of tax and welfare reform in recent years." He is concerned with why there has been so little reform and offers six lessons from history, which are mostly prescriptions for what not to do next time. Professor Lampman is more positive, emphasizing the big changes that have occurred in the last several years. But he sees the changes as raising new problems and sounds anything but optimistic about the future.

History is only "dismal" if one expected more rapid progress. Personnally I have been struck with how far we have moved, both intellectually and politically, toward a workable income maintenance system since the problem surfaced in the mid-1960s. Let us go back to 1966, which was after all only eight years ago. Looking around in that year, one would have seen a creaky welfare system, designed thirty years before for very different problems, coming under near universal attack. The AFDC program, which had been designed to handle the "temporary" problem of widows and orphans not yet covered by social security, was growing rapidly and unexplainably. Families on welfare were subject to 100 per cent tax rates, almost no aid was available for families with a male head, and the strictness of the welfare categories was maintained by a man-in-the-house searches.

The academic economists had diagnosed the problem and come up with a neat solution. They wanted to replace the whole welfare system with a negative income tax, which would guarantee everyone a minimum income based on family size and preserve incentives to work by reducing the benefit payment substantially less than one dollar for every dollar earned. The negative tax seemed to solve the problem of poverty, work incentives, and family break-up all at once. It was a clean, attractive, utopian scheme and most us, like Professor Tobin, "had some hope of seeing it adopted, but not very much." We thought there would be plenty of time to design, carry out, and analyze a negative income tax experiment before serious consideration need be given to drafting legislation.

President Johnson was not at all interested in the negative income tax. I don't think he ever explicitly rejected it; he just did not think that anything with so little appeal as welfare reform was worth thinking about. Then came the 1968 election. Those of us who had tried and failed to sell a Democratic Administration the basic idea of welfare reform assumed the jig was up, at least for a while.

But we were wrong. In the next four years events moved much faster than any of us thought possible. A Republican President proposed a basic welfare reform which looked very much like a negative income tax, and the Congress took it seriously. Indeed, considering that it was a new idea stemming from a President of the opposite party the Congress gave the plan a remarkably warm reception. The Family Assistance Plan passed the House of Representatives twice and could have passed the Senate had an ambivalent President not changed his mind in the middle of a reelection campaign. Perhaps in the heat of the election he realized that Johnson was right: there are no votes in welfare.

Professor Tobin professes not to know "whose fault it is that FAP never got through the Senate." In my opinion, although the arch radicals and the rabid conservatives deserve their share, the blame lies squarely with the Chief Executive for backing away at the crucial moment from the workable compromise worked out by Senator Ribicoff and then Secretary of HEW Elliott Richardson.

Despite these reverses, as Professor Lampman points out, the last several years have seen substantial steps in the direction of a universal income maintenance system which would eliminate poverty without discouraging work. The AFDC program has been liberalized. The Supplemental Security Income Program (SSI) is essentially a negative income tax for the aged. The Food Stamp Program now has universal federal standards and has grown into a kind of negative income tax in-kind, available to the working poor as well as to people in the welfare categories.

My own view of history is somewhat Tolstoyan -- great battles won or lost because a single soldier picks up the flag and runs the right (or the wrong) way at the crucial moment. If President Nixon had supported the Richardson-Ribicoff compromise and the Family Assistance Plan had become law, perhaps Professor Tobin would be remarking with surprise on the rapidity of progress.

Perhaps, however, he would not have been optimistic even if FAP had passed, since he saw FAP as incrementalist,"not a strategy which would lead gradually to a more fundamental reform." He believes that true progress in the income maintenance area must involve reform of the tax system and that both transfers and taxes must be handled under a single system by the Internal Revenue Service.

I disagree. I believe it is possible to have a dignified well-run income maintenance system administered by an agency other than the IRS. The new SSI system, administered by Social Security, does not seem to be obviously inferior to a negative income tax for the aged administered by IRS. The administrative problem of running an income transfer system for low income people is

quite different from that of collecting a positive tax. The accounting period has to be shorter, different kinds of information have to be collected, the definition of income may have to be different. Hence, forms and procedures will have to be different for negative than for positive taxpayers even if the same agency administers both programs.

Indeed, I would suggest that recent history may yield a seventh lesson; namely, that the strict tax approach to income maintenance has almost zero political appeal. To be sure, its proponents have not explained it adequately (even to presidential candidates) and should try harder. Nevertheless, it must be recognized that enthusiam for coupling the positive and negative tax systems remains low, especially in the corridors of the Internal Revenue Service.

But the most important lesson of the recent history seems to me that economists and other policy analysts simply have to work harder on policy problems if they are to come up with practical solutions. Solving the income maintenance problem will require more than coming up with neat sounding proposals. It will be necessary to think these proposals through carefully, to explore how they would relate to existing programs and how they would be administered. We are all a lot wiser now than we were in 1966. We know a lot about messy things like the problem of cumulative marginal tax rates and the crucial importance of accounting periods. We are all to aware of the equity problems created by the fact that any program which relates "need" to family size results in substantial transfers from small to large families with the same income level.

I am not saying that the policy analysts of 1966 were politically naive -- worse than that, we were technically naive. We were like theoretical physicists trying to build a bridge or a bomb. We simply did not understand how complicated practical problems were. Now that we do, perhaps progress will be faster.

THE CHANGING SOCIOECONOMIC LOCATION OF BLACKS IN POSTWAR AMERICA

Andrew Cherlin and Robert W. Hodge

The 1950s ended on a note of tranquility at home and peace abroad. To be sure, unemployment had been rising through the Eisenhower years, but the overall trend was far from linear, having been through a series of troughs and booms. Utopia had not yet been achieved, but as we entered the 1960s a youthful President was to excite the imagination of many to believe it was in our grasp. As we entered the new decade, there were at best only distant murmurs of the turmoil and crises it would contain: escalation of an unpopular war abroad; assassinations of a President, a Presidential contender, and a Nobel laureate among others; campus disruptions; and major riots in the streets of our cities. Although events seemed disorderly at the time, from the vantage point of the 1970s it is easy to see that the conflicts of the 1960s can at least be interpreted largely as manifestations of latent issues as old as the nation itself, to wit, social and, in particular, racial equality at home and isolation vs. internationalism in foreign affairs.

There is no orderly social theory which enables us to determine the conditions under which latent tensions become manifest conflicts. Open conflict is, however, like a sore on the social body and it inevitably brings forth efforts to ameliorate the conditions which gave rise to it. In itself, manifest conflict is neither good nor bad; like adultery it is only good or bad within a particular scheme of values. As with diseases, we cannot predict too well when social conflict will besiege us. Similarly, modest conflicts, like minor symptoms, may well unravel major ills which can be corrected before their consequences are fatal.

There is scant doubt that a rising tide of civil rights protest in the early 1960s was followed in the Johnson years by the most significant efforts to secure the equal rights of citizens through federal legislation since the Civil War. We cannot, of course, be certain that the latent racial tensions heated up by the civil rights movement were the proximate cause of subsequent Congressional activity. Similarly, we are at a loss to identify why a significant civil rights movement came to fruition in the 1960s rather than decades before. What we can do, however, is examine the changing socioeconomic location of blacks through the decade. Such an endeavor does not enable us to pinpoint the specific causes of any observed changes, but it does enable us to assess in a global way the consequences of a turbulent decade upon the relative socioeconomic position of our black citizens.

POSTWAR TRENDS IN THE SOCIOECONOMIC STATUS OF BLACKS

The absolute level of living experienced by blacks has been rising for some time, for as real income has risen in our expanding economy everyone's standard of living has improved, even those groups whose share in the total dividends of our society is less than their relative size. There is now appreciable evidence that through the 1960s the socioeconomic circumstances of blacks improved both absolutely and relatively (Farley and Hermalin, 1972; Freeman, 1973; Wattenberg and Scammon, 1973). Some of the evidence on this matter is summarized graphically in Figure 1, which shows for the postwar period the time path of (1) the ratio of nonwhite to white median family income and (2) the index of occupational dissimilarity between employed whites and nonwhites.[1] The indices of dissimilarity were computed over the major occupation groups of the U.S. Bureau of the Census; the index values represent the percentage of blacks (or of whites) who would have to shift their major occupational group in order to effect equality in the occupational distributions of the two races.

As can be seen from Figure 1, the relative family income of nonwhites is fairly stable through the 1950s at a figure just over half that of white median family income. However, through the 1960s, nonwhite family income expanded more rapidly than white family income so that it now stands at over three-fifths the level enjoyed by whites. Whether it was the result of civil rights activity, of equal employment opportunity legislation, of the poverty program, or some other innovation of the 1960s we cannot say, but evidently there were forces at work in the past decade, not operative in the more distant past, which led to a relative improvement in the economic position of black families.

The indices of occupational dissimilarity reveal a decline in the level of occupational segregation which parallels the rise in the relative family income of blacks. Thus, relative income gains were fortified by movements in the direction of occupational equality. Unfortunately, we are unable to create annual series, disaggregated by sex, measuring the relative occupational position of blacks. Somewhat more piecemeal data from scattered time points was, however, assembled by Farley and Hermalin (1972, p. 363). Their figures show that both the occupational segregation of employed black men from white males and the occupational segregation of employed black women from employed white females evidenced little change through the 1950s, but declined in the 1960s. Relatively greater gains were made by black women vis á vis white women than by black men relative to their white counterparts.

All of the foregoing results are, of course, based on gross differences between whites and nonwhites. Only a fraction of the gross differences between blacks and whites in economic and occupational outcomes can be traced to racial inequalities in economic and occupational opportunities. The gross differences conceal important differences between the races not only in access to opportunities, but also in attainment. Blacks, for example, are concentrated in the South where incomes are lower and have levels of educational attainment inferior on the average to those of whites. These factors, among others,

suffice to explain part of the observed gross differences between whites and black in both occupational and economic outcomes. It also happens that blacks have been moving toward educational equality with whites during the period under consideration (cf. Farley and Hermalin, p. 364), while assuming at the same time a population distribution by region which increasingly resembles that of whites. Consequently, it is just possible that the relatively improving occupational and economic circumstances of blacks may be traced, among other things, to their regional redistribution and improving educational attainment relative to whites, rather than to any substantial alteration of their occupational and educational opportunities during the turbulent 1960s.

MULTIPLE CLASSIFICATION ANALYSIS OF INCOME

In order to assess the possiblities set forth above, one needs to move from the analysis of gross racial differences in socioeconomic outcomes to the examination of net differences in income and occupational status, adjusted for such salient correlates of income and occupation as age, education, and region of residence. To do this, one needs to specify a model of income determination at the individual level and estimate this model in successive time periods. Published tabulations from the 1960 and 1970 Censuses of Population enable us to estimate a reasonable approximation to such a model at the beginning and end of the 1960s.

For the male experienced civilian labor force aged 25-64 in 1960, we estimated the following model:

$$Y_i = \bar{Y} + \sum_{j=1}^{2} \alpha_j R_{ji} + \sum_{k=1}^{5} \beta_k E_{ki} + \sum_{r=1}^{11} \gamma_r P_{ri} + \sum_{s=1}^{3} \lambda_s A_{si} + \sum_{t=1}^{2} \pi_t G_{ti} + \epsilon_i, \qquad \text{(Eq. 1)}$$

$$\sum_{J=1}^{2} R_{ji} = \sum_{k=1}^{5} E_{ki} = \sum_{r=1}^{11} P_{ri} = \sum_{s=1}^{3} A_{si} = \sum_{t=1}^{2} G_{ti} = 1, \text{ for all } i, \qquad \text{(Eq. 2)}$$

and

$$\sum_{j=1}^{2} \alpha_j \bar{R}_j = \sum_{k=1}^{5} \beta_k \bar{E}_k = \sum_{r=1}^{11} \gamma_r \bar{P}_r = \sum_{s=1}^{3} \lambda_s \bar{A}_s = \sum_{t=1}^{2} \pi_t \bar{G}_t = 0, \qquad \text{(Eq. 3)}$$

where (1) Y_i is the 1959 income of the ith person, (2) $\bar{Y}$ is the mean of the Y_i's, (3) the R_{ji}'s, E_{ki}'s, P_{ri}'s, A_{si}'s, and G_{ti}'s are sets of dummy variables taking on the values zero and one which describe, respectively, a respondent's race, years of school completed, major occupational group, age, and region of residence, (4) the $\bar{R}_j$'s, $\bar{E}_k$'s, $\bar{P}_r$'s, $\bar{A}_s$'s, and $\bar{G}_t$'s are the means of the respective dummy variables and, hence, are equivalent, respectively, to the proportions of the studied sample belonging to the jth race, kth educational category, rth major occupation group, sth age category, and tth region, (5) ϵ_i is a random variable with mean zero, and (6) the α_j's, β_k's, γ_r's, λ_s's, and π_t's are the coefficients we wish to estimate.

There is nothing particularly novel about this model; it is stated in the general form of a multiple classification analysis (Melichar, 1965) wherein the 1959 income of the ith person is treated as an additive function of his race, years of school completed, major occupational group, age, and region of residence. Equation 1 merely states this model; Equation 2 sets forth five logical identities which exist between the predictor variables, and Equation 3 sets forth the identifying restrictions which enable one to estimate the model, given the identities stated in Equation 2. The model is wholly analogous to dummy variable analysis (Suits, 1957) save that the coefficients associated with the predictor categories appear as net deviations from the grand mean rather than as net deviations from an implicitly omitted category.

A model nearly identical to that set forth above was estimated with 1969 income as the dependent variable for the male experienced civilian labor force aged 25-64 in 1970. There are three minor differences between the two models, save from whatever differences are incurred by the relative quality of the two censuses. First, the 1970 tabulations enable one to contrast Negroes and whites, while the 1960 results pertain to whites and nonwhites. Second, while the major occupational categories remain identical in title (after combining "operatives, except transport" with "transport equipment operatives" in 1970 to form "operatives and kindred workers" and similarly combining "service workers, except private household" with "private household workers" in 1970 to form "service workers") there are nevertheless changes in the job content of the major occupational groups between the two censuses. Finally, all 1970 occupational returns were allocated, so an "occupation not reported" category appears in the 1960 analysis, but not in the one for 1970. We doubt if these modest changes are likely to alter appreciably any major differences observed between the two analyses.

The results of the multiple classification analyses are shown in Table 1, which also gives the coefficients observed in the analysis of 1959 income in terms of estimated 1969 dollars. The transformation of the 1959 results into estimated 1969 dollars was achieved by inflating them by the ratio of the consumer price index in 1959 to that for 1969. If one compares the 1959 results when expressed in 1969 dollars with those observed for the later year, it is evident that overall pattern of results is much the same for the two periods. To be sure, 1959 coefffficients, even when adjusted to 1969 dollars, are not precisely the same as those derived from the 1970 census data. However, many of these differences are to be explained by the simple principle that if the real income of two groups rises by a common proportion (rather than a fixed amount) the absolute difference in their income levels will likewise increase. For example, there appears to be some widening of regional income differences between 1959 and 1969, but it turns out this widening is more apparent than real since their relative incomes were in fact converging slightly. We can read from Table 1 that net income of Southerners, adjusted for race, age, education, and occupation, was \$6586 (= \$7354 - \$768) in 1959, expressed in 1969 dollars. The corresponding figure in 1969 was \$8757 (= \$9579 - \$822) or roughly a (100) (\$8757 - \$6586)/(\$6586) = 32.9 percent rise in purchasing power. Similar calaculations for those living in the North and West reveals a roughly estimated rise in purchasing power of 29.6 percent. Thus, while the absolute income differential between the North and South, adjusted for race, age, education, and occupation, was increasing, their relative incomes were for all practical purposes stable.

Not all of the differences between the 1959 coefficients, as expressed in 1969 dollars, and those for 1969 can be understood by the fact that proportional increases in real income also increase absolutely the between group variance in mean levels. Among the more notable changes which cannot be explained in this way is the fact that the net income of high school graduates, adjusted for race, region, age, and occupation, slips from nearly two hundred dollars above the grand mean to more than two hundred dollars below it a decade later. A similar slippage of lesser magnitude is observed in the relative income position of craftsmen. In addition, the coefficients for the age categories suggest, albeit weakly, the opportunity advantages being experienced by the small Depression cohort as it moves through its life cycle (cf. Winsborough, 1972). Doubtless one might tease other small changes from the results, but to do so would only further distract us from our primary focus on racial differences.

THE INCOME POSITION OF BLACKS

One can compute from the results displayed in Table 1 that the net difference between the income of whites and nonwhites,[2] adjusted for occupation, education, age, and region of residence, stood at \$1213 in 1959. Converting this figure to 1969 dollars yields a racial income gap of \$1536, which may be compared with the net differential of \$1671 observed in 1969. Thus, in so far as we can ascertain, the absolute net income differential between black and white males in the experienced civilian labor force, aged 25-64, was expanding slightly through the 1960s. The observed shift appears to us, however, as well within the bounds of sampling and judgmental error, particularly in the selection of price indices to inflate the 1959 coefficients. For all practical purposes, then, the absolute net differential in the purchasing power of blacks and whites was about the same in 1969 as a decade earlier.

Paradoxical as it may seem, while the absolute net income differential between blacks and whites was, if anything, expanding, the relative income position of black males in the experienced civilian labor force, aged 25-64, was improving. We already know from the results presented in Figure 1 that gross, unadjusted nonwhite family income was improving relative to that of white family income. The present claim, however, refers to the relative income of black males in the experienced civilian labor force aged 25-64, adjusted for occupation, education, age, and region of residence. Nonetheless, the point is easy to see: given that the absolute net income differential was close to constant and the real incomes of both blacks and whites rose, the relative income position of blacks, adjusted for the variables considered here, must necessarily rise. The easiest way to see this phenomenon statistically is to express the adjusted income differential between the races as a fraction of average black income. In 1959, the adjusted income differential between white and nonwhite males aged 25-64 and in the experienced civilian labor force amounted to 35 percent of average income of nonwhite black males in the same age and employment categories. The corresponding figure for 1969, which contrasts whites and blacks rather than whites and nonwhites is 28 percent. Thus, even after adjustment for such salient features of individual income determination as occupation, education, age, and region of residence, the income position of black males is improving relative to their white counterparts. To summarize, through the 1960s, the real income of blacks and their income relative to whites was increasing, but at the same time the absolute difference in the average purchasing power of blacks and whites was certainly not declining, and may have been expanding slightly.

ON THE COST OF BEING BLACK

There is a real temptation to regard the adjusted differences between black and white income in the two periods as a plausible indicator of the cost of being black. Were one to take this step, as Siegel (1965) does with admirable caution,[3] one could only conclude from the evidence at hand that the tax for being black was virtually constant through the 1960s.

There are, however, some perils in taking the leap which Siegel made.

First, one needs to recognize that any adjusted income differential between blacks and whites is no better than the model upon which it is based. To assert that any particular adjusted differential represents the cost of being black is tantamount to asserting that the model from which the estimate is derived is itself correctly specified. Unfortunately, in the nonexperimental sciences there is no way of determining this matter definitively. For example, we have here adjusted the racial income differential for occupation, education, age, and region of residence. Even these controls have been less than exact, since the control on age is crude and it is well known that blacks are occupationally segregated within, as well as between major occupation groups. More importantly, however, there are other variables upon which whites and blacks differ, but which the census tabulations do not permit us to control. Among these, one would surely number measured intelligence, health status, parental socioeconomic status, and various characteristics of one's family of origin, including its size and stability. Short of introducing these variables into the analysis, there is no way of knowing whether the adjusted means discussed herein would be stable. Even if these variables were controlled, which it might be possible to do by augmenting census derived covariance matrices from other sources, one could not be certain that some further unspecified variable is lurking in the background, yet to be discovered. In sum, any effort to interpret a particular set of adjusted mean differences between the income of blacks and whites as an indicator of the "cost of being black" will be hard to defend substantively.

Despite the problems noted above, it is patently clear that examination of the movement of adjusted means, such as those examined herein, is far superior to working with gross differences, since one has at least ruled out some of the more plausible explanations of the improving, relative income position of blacks, such as the increasing similarity in the educational, occupational, and regional distributions of blacks and whites. Indeed, in the light of what we now know, a plausible case could be made for accepting the figures at hand as reasonable indicators of the "cost of being black"--a phrase we interpret to mean that part of the gross racial differential in economic status which can be traced neither to average racial differentials in individual ability nor aptitude and which, therefore, must be allocated to the operation of discrimination, differential opportunities, or institutional racism. For example, Duncan (1969) provides a somewhat more complete model of income determination for deriving adjusted differences in the income levels of blacks and whites. After adjusting for the socioeconomic level of family of origin (head's education and occupation), number of siblings, measured mental ability, education, and occupation, Duncan still finds for 25-34 year old males in 1964 an adjusted racial differential in income on the order of $1200 to $1400. This figure is roughly on the same order of magnitude as the ones reported herein and suggests that the racial income gap, as adjusted here, may well be recalcitrant to downward revision on incorporation of the most obvious missing variables.

A second difficulty in interpreting adjusted racial differentials in income as indicators of the "cost of being black" flows from the fact that discrimination is multidimensional. Blacks incur "costs" at every step in the career cycle. Income is but one of many foci of discrimination. Consequently, that part of the gross racial differential in income which is washed away by factoring out, say, occupation or education may itself be a product of discriminatory practices directed not at limiting the specifically economic horizons of blacks, but their occupational and educational ones.[4]

On balance, we feel it is the better part of wisdom to disassociate adjusted racial differentials in income from a terminology involving "costs." Instead, such adjusted differentials should be treated as nothing more nor less than what they are: hypothetical calculations of what the income gap would be if the races were equated on the factors considered in making the adjustments. If there are no suppressor variables at work, i.e., factors whose inclusion would widen rather than reduce the adjusted differentials, then one can consider the adjusted differentials as an upper bound on the "costs" to blacks of discrimination with respect to the variable under investigation. In this sense, one might venture that in 1969 income discrimination "cost" blacks no more than $1700. That, however, is not a very strong statement, and one is on risky ground in going beyond it.

TOWARD INCOME EQUALITY OF THE RACES

Among the many summary conclusions ventured by Farley and Hermalin in their review of changes in racial inequality during the 1960s was the statement (1972, p. 33), "Though the progress of the '60s appears rapid in a number of respects, it does not, in our opinion, presage a short run end to racial differences in income, occupation, or education." There is no denying that the relative socioeconomic location of blacks was improving through the 1960s; there is also no denying that these relative gains represent a significant change in the American social order for the simple reason that the record of the more distant past is one of absolute gains for everyone and scant, if any, relative gains for the black subpopulation. Nevertheless, these real changes hold forth scant hope, as Farley and Hermalin surmise, of any early demise in racial inequality.

This point can be developed a little more formally than Farley and Hermalin present it.

Let $\bar{N}_t$ and $\bar{W}_t$ be the average incomes of blacks and whites, respectively, in year t; $\bar{N}_{t+10}$ and $\bar{W}_{t+10}$, the corresponding means a decade later; k, the annual rate of inflation; r_n, the annual rate of real income growth for blacks, and r_w, the annual rate of real income growth for whites. With these definitions, we have the following identities:

$$\bar{N}_t\ (1 + k)^{10}(1 + r_n)^{10} = \bar{N}_{t+10}, \qquad \text{(Eq. 5)}$$

and

$$\bar{W}_t(1 + k)^{10}(1 + r_w)^{10} = \bar{W}_{t+10}. \qquad \text{(Eq. 6)}$$

Taking the ratio of these equations and setting $x = (1 + r_n)^{10}/(1 + r_w)^{10}$, we have

$$x = (\bar{W}_t\bar{N}_{t+10})/(\bar{N}_t\bar{W}_{t+10}). \qquad \text{(Eq. 7)}$$

With $z = x^{(1/10)}$, the equation,

$$(\bar{N}_{t+10})/(\bar{W}_{t+10})z^y = 1, \qquad \text{(Eq. 8)}$$

may be solved for the number of years (=y) it will take for racial equality of income to be reached, given the rates of real income growth, r_n and r_w, for blacks and whites.

Setting t = 1959 and using the average incomes of black and white males in 1959 and 1969, solution of the above equations reveal that, given the implicit differential real income growth of blacks and whites through the 1960s, racial equality in income, at least for males aged 25-64 in the experienced civilian labor force, would be achieved in 2015. Working with the income figures adjusted for age, region of residence, education, and occupation yields a virtually identical estimate. Thus, were the relative real income growth of blacks and whites observed in the 1960s to persist into the indefinite future, black males in the experienced civilian labor force aged 25-64 would achieve income equality with their white counterparts about 150 years after Emancipation.

Barring any appreciable upward drift in the real income growth of blacks relative to that of whites, such a forecast proves, upon examination, to be very optimistic. If there is any one thing we know about the economy, it is surely that blacks are more severely affected by business cycles than whites (cf., Hodge, 1973), the colloquial expression for this phenomenon being "last hired, first fired." We know of no basis for predicting the end of the business cycle and, consequently, it is almost dead certain that the rate of increase in real income experienced by blacks through the 1960s will be attenuated if the energy crisis or any other factor moves us into a recession of any magnitude. Even if the experience of blacks during the 1960s was not an artifact of the Viet Nam War, the cyclical nature of economic activity will almost surely push, given current social arrangements, the date at which racial equality of income is achieved well beyond 2015.

In so far as we see, the prospect of economic equality for blacks lies, if it exists at all, in the very distant future. At this juncture, it seems reasonably clear that differential opportunity structures are not the primary source of social inequality at large (cf. Jencks, et al., 1972). If one of our societal goals is to reduce the level of social inequality, it seems fairly certain we will not be successful if national policymakers continue to dabble with equal opportunity programs, rather than formulating and enacting an effective scheme of income redistribution. For our own part, we are convinced that whether or not a policy engendering a substantial redistribution of income is a reasonable one, it is the only certain way to achieve greater equality in the short run, including racial equality of income.

Effecting racial equality of income is, in fact, a modest goal compared to that of achieving a greater degree of economic equality across the board. The calculations presented in Table 1 imply, for example, that were black and white males aged 25-64 in the experienced civilian labor force in 1970 to achieve equivalent regional, age, educational, and occupational distribution, a national tax of $143 levied on every white male in the labor force and redistributed to the corresponding black population would suffice to affect income equality between the races. Were the tax levied only on those white males with incomes of $10,000 or more, it would amount to $370; restricting it to those with incomes in excess of $15,000 would put the figure at roughly $1100. These are not huge sums; even President Nixon can afford them, though you'd never guess it from his income tax returns.

No one, of course, seriously proposes that race be made a basis of taxation, though, in fact marital status and assorted other social characteristics already are. The point, however, is clear enough: we can achieve something close to racial equality of income in this country. To do so requires a program which insures that some income which now flows to whites be redirected to blacks. Since whites are such a preponderant majority, a small per capita "cost" on their part yields a large per capita "benefit" among blacks. Something less than fifty cents a day is a small, very small price to pay for the realization of something close to the American dream for our black citizens. Just shelling it out, however, will not be enough; we will also need policymakers and elected officials committed to constructing a bureaucratic structure for redistributing dollars rather than opportunities.

COMMENTS

We began this paper by observing that the major issues which divided the turbulent 1960s revolved around the older themes of equality, particularly racial equality, and isolationism. Obviously, domestic activity on the civil rights front stands as a potential cause of the observed changes in the socioeconomic location of blacks

in the postwar period. International relations and foreign wars seem, on the surface, remote from internal affairs. We would like to conclude by suggesting that, not only the civil rights movement at home, but the Viet Nam involvement abroad is an important key to understanding the modest relative advances experienced by blacks during the 1960s.

Nearly everyone agrees that the business cycle is a key factor in understanding the relative position of blacks. We find, over the period 1954-72, that the total unemployment rate had a correlation of -.50 with the ratio of nonwhite to white median family income and an association of .29 with the indices of occupational segregation reported above. These relationships are to be expected. Linking them to the Viet Nam encounter only requires one to see that the defense establishment is one potential vehicle for managing the volume of the domestic labor force. We find, in fact, over the period at hand that the ratio of armed force personnel to the total labor force is correlated -.43 with the total unemployment rate, a finding which begins to suggest how the domestic situation of blacks during the 1960s may have been affected by the confrontation in Viet Nam. These correlations are not, of course, decisive and, indeed, our efforts to incorporate them in a more complex scheme of causal relations has been unsuccessful. Given the crude quality of the indicators, the results remain suggestive. To the extent that the relative improvement in the socioeconomic location of blacks in recent years was tied to the consequences of our involvement in Viet Nam, the future of American blacks may be considerably less rosy than the experience of the 1960s suggests--and that was not all that rosy to begin with.

Acknowledgment: Many of the ideas discussed in this paper were developed under a grant from the National Science Foundation (#GS-1397, "Assimilation of Minorities into the Labor Force"). The calculations reported herein were supported by a grant from the Russell Sage Foundation. The support of these institutions is gratefully acknowledged, as is that of our colleague, Ricardo Klorman, none of whom bear any responsibility for errors this report may contain or necessarily share the views expressed herein. This paper is dedicated to K. C. Tyree, who is segregated in his own way.

FOOTNOTES

[1]The ratios of nonwhite to white median family income are taken, for 1947-71, from U.S. Bureau of the Census, 1972, Table 11, p. 34. The figure for 1972 comes from U.S. Bureau of the Census, 1973b, Table 8, p. 6. For 1947-59, the indices of occupational segregation were compiled from various government publications. The values for subsequent years were computed from U.S. Department of Labor, 1973, Table A-11, p. 141, and Table A-12, p. 143. The indices of occupational segregation are afflicted by two major incomparabilities. The figures for 1960 and later years include Alaska and Hawaii and are based on employed persons aged 16 and over, while the earlier figures exclude Alaska and Hawaii and are based on employed persons aged 14 and over. The figures for 1971 and 1972 are not exactly comparable to those for earlier years owing to changes in the definition of the major occupational groups over which the indices are computed; other, more modest changes of this sort afflict the series at other points.

[2]The analysis for 1959 is based on the contrast of whites and nonwhites, while that for 1969 compares whites with blacks. Consequently, nonwhite, non-Negroes are included in the earlier analysis and excluded from the latter.

[3]Siegel's work is plainly the inspiration for the present paper; though portions of his paper have been superceded by Duncan's seminal work (1969), his endeavor remains one of the best informed and sophisticated analyses of the economic position of American blacks and continues to stand as a model of scientific reporting.

[4]Reflection upon this paragraph may well convince the reader that as plausible a case can be made for interpreting the gross income difference between blacks and whites as an indicator of the "cost of being black" as can be made for any alternative indicator. The name of our game is decomposition, not labelling.

REFERENCES

Duncan, Otis Dudley
1969 "Inheritance of Poverty or Inheritance of Race?" Pp. 85-110 in Daniel P. Moynihan (ed.), On Understanding Poverty New York: Basic Books.

Farley, Reynolds and Albert Hermalin
1972 "The 1960s: A Decade of Progress for Blacks?" Demography 9 (August): pp. 353-370.

Freeman, Richard B.
1973 "Changes in the Labor Market for Black Americans." Pp. 67-132 in Arthur M. Okun and George L. Perry (eds.), Brookings Papers on Economic Activity 1: 1973. Washington : The Brookings Institution.

Hodge, Robert W.
1973 "Towards a Theory of Racial Differences in Employment." Social Forces 52 (September): pp. 16-31.

Jencks, Christopher and Marshall Smith, Henry Acland, Mary Jo Bane, David Cohen, Herbert Gintis, Barbara Heyns, and Stephan Michelson
1972 Inequality: A Reassessment of the Effect

and Schooling in America. New York: Basic Books.

Melichar, Emanuel
1965 "Least Squares Analysis of Economic Survey Data." Pp. 373-385 in 1965 Proceedings of the Business and Economics Statistics Section, American Statistical Association.

Siegal, Paul M.
1965 "On the Cost of Being a Negro." Sociological Inquiry 35 (winter): pp. 41-57.

Suits, Daniel B.
1957 "The Use of Dummy Variables in Regression Equations." Journal of the American Statistical Association 52 (Dec.): pp. 548-551.

U.S. Bureau of the Census
1963 Census of Population: 1960 Subject Reports, Final Report PC (2)-7B, Occupation by Earnings and Education.

U.S. Bureau of the Census
1972 "Money Income in 1971 of Families and Persons in the United States." Current Population Reports, Series P-60, Number 85 (Dec.).

U.S. Bureau of the Census
1973a Census of Population: 1970 Subject Reports, Final Report PC(2)-8B, Earnings by Occupation and Education.

U.S. Bureau of the Census
1973b "Money Income in 1972 of Families and Persons in the United States." Current Population Reports, Series P-60, Number 87 (June).

U.S. Department of Labor
1973 1973 Manpower Report to the President. Washington, D.C.: U.S. Government Printing Office.

Wattenberg, Ben J. and Richard M. Scammon
1973 "Black Progress and Liberal Rhetoric." Commentary 55 (April): pp. 35-44.

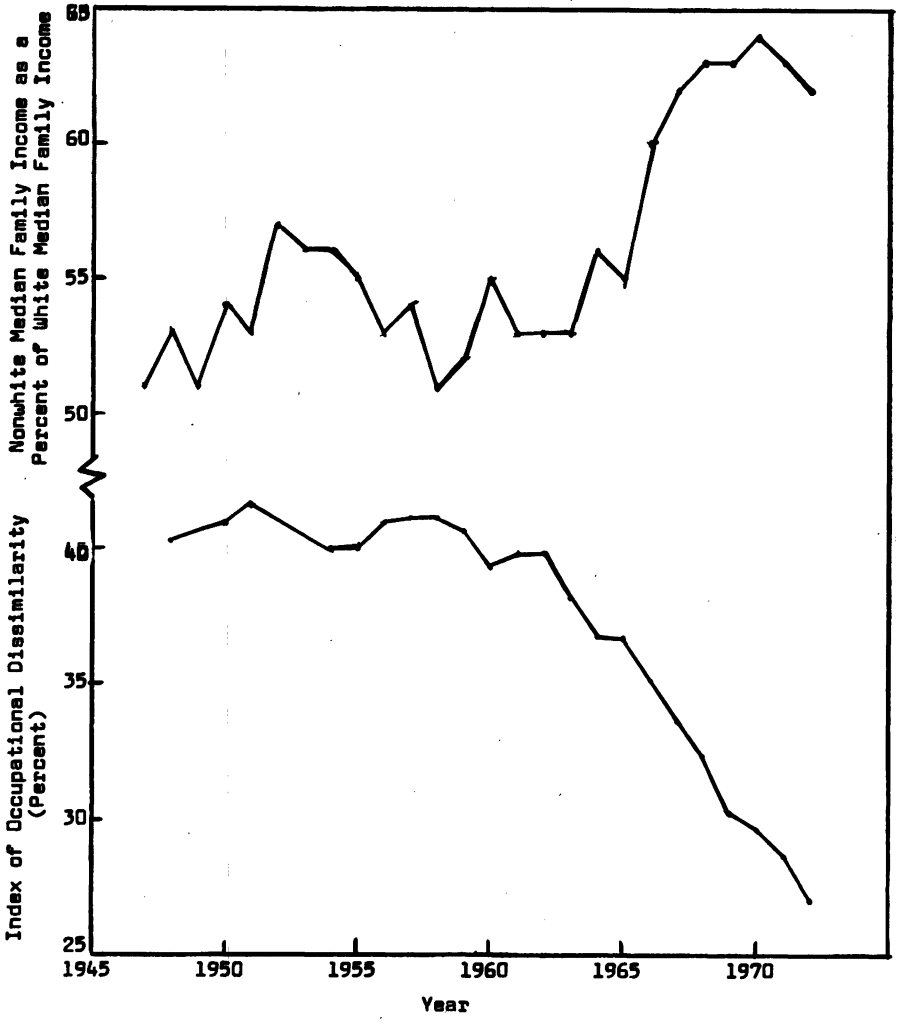


FIGURE 1. TRENDS IN THE RELATIVE FAMILY INCOME AND OCCUPATIONAL SEGREGATION OF WHITES AND NONWHITES, 1947-1972.

TABLE ONE - MULTIPLE CLASSIFICATION ANALYSES OF 1959 AND 1969 INCOMES OF MALES, AGE 25-64, IN THE EXPERIENCED CIVILIAN LABOR FORCE.

CATEGORY	1959	IN 1969 DOLLARS 1959	1969
		MEANS	
TOTAL	5847[2]	7354	9579[3]
White	6112	7687	9920
Non-White[1]	3260	4108	5936
	ADJUSTED DEVIATIONS FROM GRAND MEAN		
Race			
White	113	142	143
Non-White	-1108	-1394	-1528
Years of School Completed			
0 - 8 Years Elementary School	-1183	-1488	-2112
1 - 3 Years High School	373	469	-1039
4 Years High School	146	184	- 215
1 - 3 Years College	770	968	668
4 or more Years College	3401	4278	4100
Age			
25 - 34	-908	-1142	-1506
35 - 54	367	462	703
55 - 64	280	352	279
Occupation			
Professional, Technical & Kindred Workers	830	1044	1261
Managers and Administrators, except farm	2834	3564	3098
Sales Workers	592	745	1088
Clerical and Kindred Workers	-701	-882	-1150
Craftsmen and Kindred Workers	28	35	-142
Operatives	-502	-631	-993
Laborers, except farm	-1180	-1484	-1754
Farmers and Farm Managers	-1942	-2443	-2273
Farm Laborers & Farm Foremen	-2547	-3203	-3739
Service Workers	-1290	-1622	-2141
Occupation not Reported			
Region			
North and West	238	299	341
South	-611	-768	-822

[1] In the 1960 Census the category "non-white" is used. In the 1970 Census the category "Negro" is used.

[2] Grand Mean for all males in U.S. Bureau of the Census (1963), Table I.

[3] Weighted mean of the total white and total Negro means from U.S. Bureau of the Census (1973a), Tables I & II.

SOURCES: U.S. Bureau of the Census (1963, 1973a).

MICROANALYTIC SIMULATION AND THE DEVELOPMENT OF SOCIAL ACCOUNTS*

Steven B. Caldwell, Guy H. Orcutt
The Urban Institute

I. INTRODUCTION

The social accounting "movement" has spawned considerable strategy debate, much of it ranged across two competing positions characterized as "theorist" vs "inductivist":

> To sum up, if not caricature, the two positions: The "theorist" says, "Let us think long and hard about what we want to measure and why. Then we will feel confident about what ought to be done by way of making observations." The "inductivist" responds, "Let us see if we can measure something, for whatever reason, and standardize our measurements so that we achieve an acceptable level of reliability. Then let us study how the quantity being measured behaves." (Duncan, 1969, p. 9)

In the same article, Duncan recommended an "inductivist" priority for social reporting efforts to be directed at the measurement of social change by means of replicating important "baseline" studies, such as the 1962 O.C.G. survey, Project Talent, the 1965 E.E.O.S., et al. An "inductivist" bias can be a necessary corrective to the measurement paralysis which often stems from preoccupation with theoretical frameworks. For example, insofar as certain "theory"-oriented definitions of social indicators as components of models (Land, 1971; Wilcox and Brooks, 1971; Anderson, 1973) are understood to mean that only those variables included in precisely specified models deserve attention and measurement, the price may be a loss in that richness of the data base which often leads to improved modeling efforts.

Nevertheless, these allowances having been made, it remains true that a role in a system of relationships enormously enhances the usefulness of a social indicator. Although initially simply measuring a criterion variable may be of interest, such measurement usually stimulates interest in influencing the indicator for the better, i.e., in policy guidance. And to raise the question of the effect of various policies on a criterion almost always leads beyond the simple policy-to-criterion relationship. Intervening contingencies more proximate in time or in the causal chain (usually both) to the policy instruments must be included in the model. Some of these contingencies, because of their effects on the initial indicator, may themselves gain meaning as social indicators. Policy analyses must also take account of unintended consequences to other criterion variables, whether flowing directly from the policies or indirectly from one or more intervening contingencies. As models specifying the relationships among indicators of states of social systems are improved to take account of these complexities, the indicators themselves assume more meaning and offer more useful policy guidance. It is towards such social accounting and policy guidance purposes that our modeling efforts have been directed.

However, the purpose of this paper is not to simply urge more inductive modeling efforts but rather to suggest the advantage of a particular modeling paradigm. This approach - a dynamic, microanalytic model of family and person behavior coupled with an auxiliary model of the national economy - has been developed in a series of papers by Orcutt (1957, 1960, 1968). A partial realization of the approach constructed in the late 1950s was described in Orcutt, et al. (1961). The suggestion to apply this modeling paradigm to social accounting was first broached by Sprehe and Michielutte (1969). Our purpose is to further emphasize the advantages of the microsimulation paradigm for social accounting purposes.

II. THE URBAN INSTITUTE MICROANALYTIC MODEL

A full description of the model and of certain experiments performed with it is given in a forthcoming volume (Orcutt, et. al., forthcoming). Only a brief description, drawn from the volume, is given below.

The Urban Institute microanalytic model of the U.S. population takes a sample representation of the U.S. population at some point in time and modifies the sample in ways which simulate the behavior of individuals and families in the population over a year's time. The result is a new sample representation of the population one year later. Both the number and characteristics of persons and families in the sample are changed. This new sample can then be modified by the model a second time, and so on.

The sample population, moved forward in time by applying the microanalytic household model, is comprised of three types of entities: persons embedded within nuclear families which in turn are embedded within interview units. For purposes of the paper, a nuclear family consists of either an unmarried person or a couple with own children, if any, still living at home. An interview unit consists of one or more nuclear families, related by blood, marriage or adoption and residing together.

---

*Several government agencies and foundations have financed the model-building work described in this paper. Four have played critical roles - the Office of Economic Opportunity, the National Science Foundation, the Social Security Administration and the Ford Foundation. Opinions expressed are those of the authors and do not necessarily represent the views of the Urban Institute or its sponsors.

The microanalytic model is composed of an interrelated set of processes, or operating characteristics, each of which embodies a set of behavioral or accounting relationships that specify, for each entity, the outputs generated by that entity given inputs into the entity and its previous state. The current microanalytic model consists of four broad groups of operating characteristics: (1) demographic, including leaving home, divorce, birth, death, aging, marriage, education and geographic mobility; (2) labor, including labor force participating, occupation, hours in the labor force, unemployment, hours worked and wage rate; (3) taxes and transfers, including private intergenerational transfers; and (4) income and wealth of nuclear families including nuclear family consumption, wealth accumulation and income from wealth. Work on, and implementation of, the above last two groups of operating characteristics is still at an early stage of development.

Running along with the micro model is a small macro model which is linked to the micro model by means of a time series data bank. The primary function of the macro model is to link macro variables, such as GNP and the GNP price deflator, as well as micro variables, such as labor force participation and unemployment of persons, to macro-policy actions as embodied in the realized level of unemployment. Data from the time series data bank can be used to influence how the micro model updates information about the sample and can, in turn, be modified by summary information drawn from the micro population data arrays. For example, the total number of persons simulated to be in the micro population can be stored in the time series data bank each year. The macro model can also send information to and receive information from the time series data bank. This entire computerized structure is called the Microanalytic Simulation of Households system (MASH).

III. APPLYING THE MODEL

Experiments with the model can be performed to explore the effects over time of particular alterations in one or more initial conditions, specifications or parameters. These experiments may be intended to simulate the impact of either a deliberate public policy or an ongoing socioeconomic trend outside the control of public policy. Experiments are performed as follows. First the behavior of the population is simulated for the years of interest with the "standard" set of operating characteristics. The simulation is then repeated over the identical period beginning from the same initial population but with certain alterations in the model. Any differences in output between the two runs are, within certain confidence intervals dictated by sampling variability, attributed to the alterations. In this way predictions of the effects of a range of potential policies can be made as a partial basis for policy selection. Of course, the accuracy of the predictions depends entirely on the validity and scope of the understanding built into the model.

Since samples of microunits are used to represent corresponding populations (of families, sub-families and persons), the range of potential output from an experiment is limited only by the scope of the variables describing the microunits. At any stage in the simulation the sample of families and individuals can be read into a compatible package of analysis routines. Output from "control" and "experiment" simulations can thus be compared in great detail.

The ability to perform experiments using the model greatly increases its usefulness for social accounting. For most social and economic indicators, some, if not all, of their claim on our attention is due to their impact on other variables. This is most obviously the case with macroeconomic indicators, whose importance rests ultimately on their presumed connection, however indirect, with the welfare of individuals. But it is just as true for micro-level indicators. They also derive at least some of their importance from their presumed effects on other indicators. Thus, to specify the effects of being in a particular status is to clarify the importance of indicators of that status. An example is the social indicator 'marital status'. Whether, if considered somehow in isolation, differences in marital status imply differences in individual welfare is problematic. But to the extent systematic relationships between, say, age at first marriage, and longevity, income, fertility, occupational mobility and other variables are uncovered, then to that extent marital status also becomes worthy of attention as a social indicator. A social accounting framework must incorporate the linkages between indicators, so that the implications of changes in particular indicators will be apparent.

To illustrate the use of the microsimulation model to explore the effects of changes in a given social indicator, we present selected results from an experiment performed on the model. The experiment was conducted to explore the implications of a sharp change in the pattern of age at first marriage. Although it is certainly possible to conceive of public policies having effects (probably unintended) on first marriage patterns, it is more appropriate to view the experiment as simulating a social change (change in a social indicator) over which public policy may have little control but whose consequences such policy may have to take into account. Since such social changes are continually modifying the policy environment, alleviating some problems, worsening others and creating entirely new ones, it is just as important for the purposes of policy analysis to be able to predict the consequences of nondeliberate, as of induced, social changes (i.e., changes in social indicators).

Both the "control" and "treatment" used an initial population drawn from the 1960 census and both generated ten years of simulated behavior. In the treatment the equation generating first marriage probabilities was altered so that all never-married persons were assigned a first marriage probability equivalent to the

one assigned to a person four years younger in the control. The effect of this alteration is to shift the age pattern of first marriage four years to the right, while preserving differentials by age, race, sex, education and marital status. Table 1 presents selected differences between control and treatment. Significant differences were found for the number of marriage, divorces and children involved in divorce, births, and deaths, and also for the total persons in the labor force, total employment and gross national product. The results depend, of course, on the validity of the causal assumptions embedded in the model.

TABLE 1

FOUR YEAR DELAY IN AGE AT FIRST MARRIAGE EXPERIMENT

$$[\%\ \text{change in } X = \frac{\text{experiment-control}}{\text{control}} = a_o + a_1\ (\text{YEAR} - 1960)]$$

X	a_o	a_1	Total Impact over the entire period 1960-70
Marriages			
(t-statistic)	-43.73	2.27	-5,851,000
	(34.6)	(10.6)	
Divorces	-3.18	-2.65	-784,000
	(0.3)	(10.0)	
Births	-5.62	-3.02	-6,882,000
	(1.1)	(8.2)	
Deaths	-0.42	-0.62	-590,000
	(0.4)	(7.9)	
Total Population	0.20	-0.39	-6,292,000
	(8.1)	(33.2)	
Children in Divorce	0.31	-2.62	-832,000
	(2.8)	(15.7)	
Female Headed Families	-8.57	0.90	*
	(1.5)	(.9)	
Total Persons in Labor Force Any Time	-0.64	0.15	296,000
	(7.4)	(8.8)	
Total Employment	-0.64	0.14	250,000
	(7.6)	(8.6)	
Average Earnings of Full-Time Black Males	-2.80	1.44	*
	(1.3)	(2.8)	
GNP	-0.61	0.01	-$39.9 billion
	(5.6)	(0.4)	

*Impact considered not significant if t-statistics for both slope and intercept are <3.0.

Experiments such as this in effect explore the importance of various social indicators. The age-at-first-marriage indicator gains importance, for example, insofar as age-at-first-marriage has an effect on such variables as employment, fertility, children involved in divorce and gross national product. Finally, insofar as we can link some public policy to a change in first marriage behavior, we can explore the secondary policy implications flowing from that linkage. In this way the model functions as a social accounting system relating public policies or social trends to changes in other related criterion variables.

IV. THE MICROSIMULATION PARADIGM FOR SOCIAL ACCOUNTING

In this section we summarize what we believe to be the major advantages of the approach described above for social accounting.

A. Microanalytic Focus. By directly representing individuals, nuclear families and families, the micro analytic model can directly treat distributional questions. Certainly questions of the distribution of benefits and costs of alternative policies are among the most important policy analysis ought to treat.[1] Moreover, micro representation enables indicators for very detailed subpopulations to be presented simultaneously with indicators for larger populations. Decisions between aggregate vs. disaggregate indices can be alleviated in certain cases by presenting indicators for several levels of aggregation at once. Micro representation also makes it possible to utilize research results concerning micro units. The far greater number of microunits than of macrounits means that problems of hypothesis testing and parameter estimation are somewhat less severe.

B. Realism of the Model. Solving the microanalytic model using simulation techniques imposes certain costs, but it also means that many fewer concessions need be made in the specification of relationships than is the usual case when models are solved analytically or using the numerical transition matrix approach. (This advantage of course follows from the solution mode, not the level of analysis.) Parsimony and elegance can be subordinated to the real needs of policy analysis - validity and predictive accuracy. The potential realism of the model is therefore greater. Analytic approaches, though preferable, are currently infeasible when nonlinearities, feedbacks, and other complexities are introduced. Transitional matrix approaches quickly become intractable as the number of dimensions describing the state of the population increases. By in effect having one function to predict change in each separate dimension, the problem of estimating an enormous number of transition probabilities as the number of categories and dimensions increases is avoided.

C. Output Capabilities. Each simulation in effect creates a simulated public use sample. This method of representing joint frequency distributions has considerable advantages as indicated by the increasing use of public use samples by the Census (Orcutt, 1973). Moreover, the model generates life history data for each micro unit, including, if applicable, characteristics of its unit of origin. Thus, suitable

[1]The particular focus of the Urban Institute model is the distribution of disposable income and of wealth among families and individuals in the U.S.

interrogation after a generation of simulation would produce data that would serve as a check on the validity of the intergenerational relationships embedded in the model. The possibilities of developing a dynamic model of inter- and intra-generational mobility are greatly increased.

D. Macro-Micro Linkages. The presence of the auxiliary macro model creates the capability of exploring distributional implications of major macroeconomic changes, including those induced by monetary and fiscal policies. Moreover, individual and family decisions affect macroeconomic variables. Currently these linkages are primitive, but the presence of both levels in the same model at least poses challenging questions of linkage between individual and national decisions.

E. Comprehensiveness. The Urban Institute model provides a framework in which to imbed a great deal of interdisciplinary research. The model functions, however, as a core model. Linkages to particular policies must be added using further evidence from evaluation studies, experiments, and analyses of non-experimental data. Given the additional links, the model can be used to explore the impacts of a particular policy on a wide range of criterion variables. The model's unusual breadth does not, of course, eliminate the need for ceteris paribus assumptions but it broadens the amount of information generated by each simulation experiment. One useful by product might be an increased sense of coherence and interrelatedness to existing knowledge.

F. Direct Focus on Change. The model is dynamic. It generates individual and family histories directly and consistently. Distributions of the population by various characteristics are the natural outcomes of the trajectories of its components. The paths by which change can be influenced are thus directly incorporated.

V. CONCLUSION

Certain advantages of the microanalytic paradigm are of course shared by other approaches. But the combination of features of the microanalytic paradigm makes it quite attractive for social accounting and related policy analyses. Work is continuing on improving the specification and estimation of the relationships in the present model. We hope the present model is a useful first step toward a major social accounting and policy analysis tool.

References

1. Anderson, J.G., (1973) "Causal Models and Social Indicators: Toward the Development of Social Systems Models." American Sociological Review, 1973, Vol. 38, June.

2. Duncan, O.D., (1969) "Toward Social Reporting: Next Steps," Russell Sage Foundation.

3. Land, K.C., (1969) "On the Definition of Social Indicators," the American Sociologist, 6, November.

4. Orcutt, Guy H., (1957) "A New Type of Socio-Economic System," Review of Economics and Statistics, May 1957.

5. _____, (1960) "Simulation of Economic Systems," American Economic Review, Vol. L., December.

6. _____, M. Greenberger, J. Korbel and A. Rivlin, (1961) Microanalysis of Socio-economic Systems: A Simulation Study. Harper and Row, New York.

7. _____, (1968) "Research Strategy in Modeling Economic Systems," in The Future of Statistics, Edited by D.G. Watts, Academic Press.

8. _____, (1973) "The Affinity of Public Use Samples and Microanalytic Models," Urban Institute Working Paper 509-3.

9. _____, Steven Caldwell, Harold Guthrie, Gary Hendricks, Gerald Peabody, James Smith, Richard Wertheimer, (1974) "Microanalytic Simulation for Policy Exploration," Urban Institute Working Paper 509-5.

10. Sprehe, J.T. and R.L. Michielutte, (1971) Final Report: Simulation of Large-Scale Social Mobility: Toward the Development of a System of Social Accounts, N.S.F. Grant No. GS-2311.

11. Wilcox, L.D. and R.M. Brooks, (1971) "Toward the Development of Social Indicators for Policy Planning," paper presented at the Annual Meeting of the Ohio Valley Sociological Society, Cleveland.

SOCIAL STATISTICS FOR PUBLIC POLICY

Robert Parke and Eleanor Bernert Sheldon
Social Science Research Council

In its discussion of the various types of users of the products of the federal statistical system, the President's Commission on Federal Statistics gave first place to policymakers. Consideration of the use of statistics by policymakers was followed, in order, by the uses by program managers, evaluators of government programs, exploratory research, industry and trade associations, state and local governments and, finally, the public.[1] This sequence of presentation accords fairly well with the priorities assigned to audiences for statistics in the budget justifications of federal statistical agencies--priorities which in turn reflect the expectations of appropriations committees.

In the same report, the Statistics Commission quoted, with evident approval, a government official, as follows: ". . . the government is simply not good at defining what it wants to do in terms of needed social science research . . . the government, in general, can only articulate the area in which it needs information."[2] The juxtaposition of these comments--priority to the data needs of policymakers, and the limited ability of government to articulate its data needs--is the basis for our topic, which has to do with the developing relationships between social statistics and public policy.

The comment of the anonymous government official must, of course, be qualified by recognition of a number of outstanding statistical programs designed to directly serve policy purposes. The unemployment statistics, the surveys of the aged population which provided much of the empirical basis for Medicare, and the statistics on poverty are a few examples of the many that come to mind, which serve directly to measure needs and to provide the factual basis for government programs to meet those needs. These are a few illustrations of social statistics in the service of policy. At their best, they represent well-planned surveys and analyses in response to an expectation of policymakers that a major program shift is in order. These types of statistical programs require, for their policy usefulness, the active participation of policymakers in determining their substantive content, since such programs are visualized as contributing to the solution of social problems. In these circumstances, the policymaker's perceptions of need are properly accorded priority in decisions to undertake the programs.

Our comments are addressed to the primacy accorded the expressed information needs of federal policymakers in shaping the social statistics enterprise generally. In our view, this set of priorities distorts the enterprise in two key ways. First, by emphasizing the expressed needs of policymakers it undervalues the potential contribution of social statistics to the definition of policy problems as well as to their solution. Problems that can be anticipated often go unrecognized until it is too late, and statistics capable of assisting in the resolution of problems often are based on an excessively narrow definition of what the problem is. Second, by emphasizing the needs of federal policymakers, the prevalent priorities undervalue the role of other constituencies for social statistics, which play important roles in the policymaking process. Such constituencies include state and local governments and substantial segments of the public at large.

These considerations lead us to suggest that the federal activity in social statistics should display more initiative in communicating the meaning of the data, through development of analytical time series and of measures which derive their significance from models of social processes, through projections of social data, especially for areas below the national level, and through social statistics chartbooks which draw upon all sources of competent data, and other means of communication. We also believe that federal statisticians should do more to cultivate their role as one key element in a complex of interrelated activities which includes the best of privately conducted social research and which serves all levels of government, private associations and businesses, and the public at large. By such efforts, statisticians can contribute to policy by assisting in the perception of problems, the anticipation of problems, and outlining the need for "adaptational" as well as "manipulative" policy strategies.

Policy Knowledge

Of policymakers, the Statistics Commission said:[3]

> "Policy-making, as the term is used by this Commission, is a set of activities including definition of problems, identification of solutions, and choice among them. The definition of a problem is a political-technical process involving translation of public perception of a situation into remedial legislation."

The Statistics Commission differentiated policymakers from program managers, whose data needs it "broadly characterized as performance data, describing resource uses and outputs of the programs administered." A related distinction is made by Biderman in his differentiation of the several levels of social information.[4] The summary is by Peter Henriot:[5]

> ". . . Biderman notes that there are three distinct uses of data which should not be confused mentally or organizationally. The lowest or most specific level of data is 'information'--data intended for use at the operational level. The next level of data is that designed for overall

administration and management purposes, and is termed 'intelligence.' The third and highest level of data is termed 'enlightenment,' and is designed for contributing to public understanding and formation of general policy . .

"An example of the various types of data outlined by Biderman would be helpful. Take the example of police records:

"--'Information,' operational data, might be the number of complaints, of routine pickups, of nightly investigations, etc.

"--'Intelligence,' administrative data, might be the number of squad cars available, of business and residences within a given area, of insurance rates on homes, of persons on probation, etc.

"--'Enlightenment,' public policy data, might be the number of criminal acts as related to characteristics of the neighborhood, the percentages of expenses spent on prevention, apprehension, and correction, etc."

Biderman's formulation views "public understanding" as a key function of statistics for policy use. For, in his words, "In a democratic social order . . . policy knowledge and public enlightenment are closely related." The same view was expressed by Otis Dudley Duncan at the 1972 meetings of this association. Our views concur with theirs. A society that relies as much as ours does on the activity of private associations and businesses, and on a high degree of participation by all levels of government and the public at large in the development of public policy, cannot afford to let its statistical agenda be limited to the perceived information needs of federal public policymakers.

Broadening Perception: Health

If the policy uses of social statistics are seen primarily as a matter of meeting the perceived information needs of administrators, we limit policy-relevant social statistics to questions raised by those whose responsibilities may cause them to overlook some important questions, and to those matters covered by the authority, competence, and traditions of administrators. No one will deny the importance of such matters. But this is no justification for letting the statistician's agenda be governed by the perceived information requirements imposed by a social engineering approach or letting it be limited to "policy-manipulable" variables--that is, those subject to the control of the agencies responsible.

Let us illustrate the distinction we are making. In 1973, a book appeared entitled Health Status Indexes. The book presents the edited results of a conference held in October 1972.[6] It appears from the papers and the published discussion that the intended audience for the book consists of decisionmakers whose chief stock in trade is the delivery of health services, and that the purpose of the work represented in the book is to devise summary measures of population health status which will assist decisionmakers in the implementation of a policy whose main outlines have been decided upon. For example, there are papers on estimating needs for health services from household survey data; papers discussing how to incorporate the prognosis characteristically associated with a condition into a measure of current health status; and studies empirically assessing the value that patients place on an impaired life, relative to the value they place on a surgical procedure with a specified probability of killing them or providing a full cure.

The fruition of these efforts should contribute substantially to rationalizing the mix of health services and their deployment. But it will contribute very little to the definition of health problems. That has been and is being accomplished in part by statistical studies of smoking and lung cancer, drinking and drug abuse, tension, and auto accidents. To a health establishment committed to dealing with those conditions that can be dealt with by providing health services, such studies as these may appear to be of limited use for decisionmaking. But some health agencies support such studies, appreciating that, if they provide little direct assistance to decision, they make an immense contribution to cognition, that is, the intellectual mapping operation that plays such a critical role in the definition of the problem. The role of statistics in assisting the redefinition of health problems is probably most spectacular in the case of smoking, partly because it led to new departures in health policy.

Defining the Problem: Manpower Policy

The 1973 Manpower Report of the President contains a chapter which represents a new departure in the attempt to define the problems with which manpower policy must cope. Entitled "Population Changes: A Challenge to Manpower Policy," the chapter describes recent national population changes stemming from the wide fluctuations in national birth rates over the past 30 years, and outlines their implications. Drawing in part on materials developed by the Population Commission, this report reviews the probable impact of these changes on the female labor force, low income and minority groups, the geographic mobility of workers, unemployment, GNP, consumption patterns, schools, health services, and other matters of policy concern. The intent is to anticipate the effects of population changes and to sensitize manpower policy to the requirements these changes impose. The editors of the Report acknowledge the major contribution to the chapter made by the Bureau of Labor Statistics.

The contribution here is to invite the attention of policymakers to developments not under their control that may alter the way they wish to conduct affairs that are under their control.

Such guidance serves policymakers by assisting in the development of what Biderman calls "adaptational strategies." Biderman constrasts adaptational strategies (that is, coping with the consequences of developments in realms the policymaker has no control over) with "manipulative strategies," which deal with matters the policymaker can control or influence. For example, our interest in weather reports is not eliminated by our inability to manipulate the weather; we need them because they "facilitate our adaptations to the phenomenon."

Social intelligence of this sort is also illustrated by statistics showing pre-Watergate declines in public expressions of confidence in government, as well as declines in the proportion of the electorate voting--a shift which reputable analysts have attributed to an increasing public sense that their vote makes little difference. Whether or not a policymaker can develop short-term responses to such developments, they are (as Otis Dudley Duncan noted at these meetings last year) surely something he ought to know about.

More on Defining the Problem: Modeling Social Processes

In speaking of the policy relevance of social statistics, we wish to emphasize its contribution to the definition of problems as distinguished from their solution. This contribution is pervasive, but is particularly clear in the case of statistics incorporated in models of social processes. For example, Featherman and Hauser have presented a cogent statement about the policy relevance of their work on path-analytic models of occupational mobility. This work, using the survey resources of the Census Bureau, will provide comprehensive trend data on the degree of inequality of opportunity in the United States and the changing importance of ethnicity, family background, education, and other factors in occupational advancement. Featherman and Hauser say:[7]

> ". . . there are five areas in which we think our study can contribute to policy formation.
>
> "(a) In the assessment of widespread beliefs about equality of opportunity and factors affecting it. We think that the 'debunking' function of OCG findings should not be underrated. For example, findings from the 1962 study cast a great deal of doubt on the utility of concepts of a 'cycle' or 'culture' of poverty, and more specifically on the suggestion that family instability was the major source of white-black achievement differentials . . .
>
> "(b) In locating and defining the problems of specific population subgroups . . .
>
> "(c) In providing an overall model of the process of social and economic achievement which can serve as a frame of reference for discussions about specific aspects of that process . . .
>
> "(d) In providing a set of current trend estimates on major features of the process of social achievement . . .
>
> "(e) In improving the measurement of processes of social and economic achievement. We think that our investment in improved measurement techniques--and also in new techniques of data reduction and analysis--will be quite as important for the policy applications of our findings as for their purely academic uses. We think our proposed innovations can contribute both to the quality and legitimacy of the information we can supply and to the development of methods for future replications and other related studies."

Halliman Winsborough has developed and presented a method for preparing annual estimates of age, period, cohort, and education effects on earnings by race.[8] It is intended to illustrate the potential uses of the Current Population Survey basic data files for social trend analysis, through models of social change. This is a use for which the CPS is uniquely adaptable given its design and size, the period it covers (since 1947), and the care with which over-time consistency of statistical practice has been ensured. Mason and Hodge have proposed a related analysis, with additional features which recognize that period effects are not uniform throughout the country, and which seek to relate the CPS data to data on political and institutional changes.[9] It would be a significant loss to social statistics if these proposals were to founder on the public unavailability of such files for the period before 1968, and the question of whether the earlier data will ever be usable.

Projecting Consequences: The Supply of Teachers

The social statistician knows something about how the society works that is capable of informing and directing the deliberations of policymakers. However, with all their competence, industry, and integrity, statisticians remain fairly passive when it comes to exploiting what they know to assist in the anticipation of policy problems. Problems that can be anticipated too often go unrecognized until it is too late.

Let us illustrate with an implausible case. Take a closed population for which high-quality current estimates of the population by age are available, and for which good current statistics on school enrollment and teaching manpower are regularly published. Should not such a population be expected to avoid major imbalances in teacher supply relative to demand? One might expect so, but that is not what has happened. Instead, we have, over the past 15 years, seen a drastic shift from a situation of undersupply to oversupply of teaching personnel. We do not refer here to the debate over whether graduate schools have

generated an oversupply of Ph.D.'s, but rather to the situation with respect to the supply of persons qualified to teach elementary and secondary school. Briefly, estimates and projections published by the National Education Association in 1971 showed that in the late 1950's the nation's teachers colleges were producing half the annual number of additional teachers needed for primary and secondary schools, in 1968 they were producing just about the number needed, and by 1972 they would be producing double the number needed.[10]

This major dislocation in the labor market deserves to be recognized as one of the most anticipatable dislocations on record. The baby boom of the 1950's and the decline in the absolute number of births starting in 1961, in combination with the rise in teachers college enrollments, made a severe oversupply of teachers in the early 1970's next to inevitable. One needed no birth projections to anticipate it; recorded births would have shown it.

We have not studied the process by which this dislocation was permitted to develop. Hence we are not in a position to state at what juncture the breakdown in foresight and planning occurred. We suspect that a significant role was played by the fact that the statistics which provided a basis for anticipating the problem were known at the national level but not at the state level at which primary and secondary school policy is made. The decline in the population under five years old was obvious in national population estimates by the mid-1960's but the state population estimates and projections of the Census Bureau contained no age detail, and no mid-decade census was available to show it. The Office of Education's projections of school enrollments and teaching graduates were published for the U.S. only, while unpublished projections for states were available on request. Thus the states were left to their own devices. Local school districts took school censuses, but they must always hedge against the substantial effects of net migration on their enrollments, and in any case they have nothing to say about teacher supply. In sum, it seems as though state policy with respect to graduating teachers generally operated on the assumption that each state could export any excess, although it would have been obvious, had anyone looked, that the sum of such implied exchanges netted out to a substantial national oversupply.

More on Projecting the Consequences: Declining Metropolitan Areas

Projecting national trends onto local areas is, of course, a highly uncertain business. No national agency, or any other, is in a position to forecast with precision which states or localities will suffer what degree of dislocation. But that some, perhaps most, areas will experience it is obvious, and it should be within the resources of a national agency to give statistical description to a few plausible projections, any one of which will yield odds that a state or locality risks substantial dislocation.

A development of this sort is in the offing with respect to the population of metropolitan areas. In the 1960's only one large metropolitan area experienced an absolute decline in population. Yet it is certain that if the national birth rate maintains its current low level or continues to drop, large numbers of metropolitan areas will experience substantial absolute declines in population in the near future.[11] This is the inevitable result of a national convergence toward zero population growth in combination with the pattern of intermetropolitan migration flows. With the drying up of nonmetropolitan sources of migration to cities, intermetropolitan migration dominates. The pattern of intermetropolitan migration is that migration is heavily focused on a few rapidly growing metropolitan areas. Most areas have maintained positive growth rates despite net outmigration, thanks to the excess of births over deaths. As that excess declines, we may expect absolute population declines in many areas.

The consequences and the policy implications of such a development are not at all clear. But they will never be clear until the prospect is addressed, and it will not be addressed until it is stated. Our own view is that, on balance, such a development is probably to be welcomed. But others may not welcome it if it takes them by surprise, and there may be real, as distinct from symbolic adjustments,which are required.

The fact is we have no experience of this particular development. We have experienced population decline in railroad towns, in agricultural marketing centers, in half the counties in the nation (mostly rural) and in several metropolitan central cities. But (the trivial decline in Pittsburgh aside), we have never experienced it in an entire large metropolitan area. What will be the consequences for land values? What will be the consequences for the amortization of fixed costs, such as roads and schools? What adjustments to decline, if any, are required? No one knows, because no one has asked. And no one will ask until the prospect is stated in numbers. Such a statement does not require that an agency estimate rates of projected decline for individual areas. But, distributions of areas by growth and decline, given plausible assumptions as to national population growth and patterns of intermetropolitan migration? Surely that would be an easy and a legitimate extension of recognized projection techniques.

Critiquing the Policy: "Balanced Growth"

We have spoken of the contribution of social statistics to the definition and the anticipation of problems addressed by policymakers. Such statistics can also play a role in critiquing positions set forth in policy discussions. Much of the discussion of "balanced growth" in recent years has been founded on the notion that the United States has too many of its people in cities, and that excessive urban growth could be countered if migrants found opportunities elsewhere, in new towns or in "growth centers," that is, communities removed from the major urban

centers, which show promise of rapid growth if deliberately stimulated by the government.

The professional skills of statisticians give them no more qualification than anyone else to comment on the judgment that we have too many people in cities. Nonetheless, statisticians are acquainted with a range of data that can be used to put this judgment in perspective. Furthermore the plausibility of a growth centers policy as a means of attentuating population growth in major urban centers can be, and was, subjected to specific examination by demographers, economists, and statisticians on the staff of the Population Commission, a body whose responsibilities included the preparation of recommendations on population policy for the United States.[12]

With respect to the judgment that we have too many people in cities, statistics were brought to bear in support of the following points:

1. The United States is a metropolitan nation. Most families live in metropolitan areas; most births, deaths, and migrations take place within or between them. Today 70 percent of the American people live in SMSA's.

2. Metropolitan population growth is a basic feature of the social and economic transformation of the United States, that is, the transition from an agrarian, to an industrial, and now to a service-oriented economy. Metropolitan growth is the form that national and regional population growth have taken.

3. The most rapid growth in the decade of the 1960's occurred not in the largest areas, but in areas with populations of between one and two million. At the same time, as metropolitan populations have grown, many central cities have been losing population.

4. The growth of metropolitan population is now mainly due to the excess of births over deaths in metropolitan areas and to immigration from abroad; not to rural-to-urban migration. A projection prepared by the Census Bureau, at the Population Commission's request, showed that if current trends continue, other parts of the United States will contribute four million migrants to the metropolitan population between now and the year 2000, while <u>immigrants</u> will add about 10 million.

The foregoing considerations suggested that the trend toward bigness, if undesirable, could not be substantially checked except as national growth is slowed or stopped. This was specifically examined by construction of a hypothetical growth-centers policy to learn how much the growth of the large metropolitan areas might be reduced if the growth of smaller, less congested places were stimulated. The results showed that if these places were stimulated to grow by 30 percent each decade from 1970 to 2000, their population might absorb about 10 million of the growth which is otherwise expected to occur in areas of one million or more, assuming the 2-child national projection. However, these large areas would still increase by 70 million persons under the same national projection.

These and related findings speak directly, and unfavorably, to the policy choice formulated by the President's National Goals Research Staff in its 1970 report:[13]

> ". . . we need to decide on whether or not we will adopt a deliberate strategy to encourage internal migration to negate the forecasts of ever-growing urban congestion in a few megalopoli. A viable option for such an alternate strategy is a policy of encouraging growth in alternate growth centers away from the large urban masses, coupled with a complementary effort of the use of new towns."

The Population Commission's findings say, in effect, that at the present stage of the nation's demographic development, there are no redistributional substitutes for lower <u>national</u> population growth rates. Whatever may be said for a "growth centers policy," the reduction of metropolitan population growth is not one of them. The policy contribution of social statistics in this instance is to point out that some policies under consideration are live options and some are dead ones. The statistical analysis does not tell us which of a variety of real-world options is preferable. What it <u>can</u> do, if heeded, is to narrow the range of consideration to those options that make sense in terms of a scientific understanding of how, in relevant respects, the world works.

<u>Communicating the Meaning of the Data</u>

The federal statistical establishment annually disgorges immense amounts of social data, only a fraction of which gets distilled as knowledge about social structure or social change. Salutary exceptions to this rule include the Census Bureau's <u>Current Population Reports</u> on changes in the social and economic conditions of blacks, Mexican-Americans, youth, and city populations. In addition, recent publications on "Age Patterns in Medical Care, Illness, and Disability," "Health Characteristics of Low-Income Persons," and other topics, have shown a recognition by the National Center for Health Statistics that its data, traditionally focused on specific health conditions and on types of health care, can be reorganized to describe important social trends.

In general, however, federal statistical reports devote a great deal of attention to how the data were put together (as they should) and very little attention to what the figures mean. This is not, of course, to suggest that statistical reports ought to contain policy comment; they should not. But, if the careful documentation of methodology is the primary responsibility of the statistician, surely a close runner-up is the responsibility to utilize accepted analytical techniques and methods of data presentation which will enable the data to tell the story that will

not be told in the absence of analysis.

An example is the information on income shares which has been published by the Census Bureau now for a quarter of a century. These are the figures showing what fraction of total personal income is received by the one-fifth of families receiving the smallest incomes. The analysis and presentation of data by income shares has a number of features of a good indicator. It is established over a long period of time with annual readings. It has been worked over by statisticians to a point where they have a pretty good idea how it acts, and why. It packs much data into few figures, and it answers a pointed question about an important aspect of our collective life.

Expectation of life is another figure with many of the same properties. An interesting feature of this figure is its seeming simplicity and its actual complexity. It is a hypothetical figure based upon the mathematical manipulation of population and mortality data. The simple statistics in this instance are the numbers of persons in the population classified by age, sex, and color, and annual numbers of deaths classified the same way. Only when the demographer or actuary processes these numbers through his mathematical model do the figures stand up and tell a story. Moreover, the message they convey is relatively unambiguous compared with such ostensibly simpler measures as the crude death rate. Expectation of life has the added virtue that it is manipulable as a component in mathematical models of the population, which renders it easily used in analyses of population growth and in population forecasting.

Possibilities for development of analytical measures may be illustrated by the prospect for an advance in knowledge about family dynamics, using available statistical resources. In a compendium of statistics on changes in the American family, Abbott Ferriss juxtaposed official statistics on divorce, classified by year of marriage, against the original number of marriages, so that one could track the divorce experience of people married in a given year.[14] Kenneth Land used Ferriss' data as the basis for development of a mathematical model describing what Land calls the "divorce trajectory" of marriage cohorts.[15] Land's work accomplishes an admirable summarization of many observations in terms of two numbers: The proportion of a marriage cohort ultimately divorcing, and the velocity at which divorces occur over the cohort's lifetime. Furthermore, it expresses the results in terms amenable to manipulation in a model of the family life cycle. Land's work on this topic addresses an area which has thus far been relatively innocent of the sophisticated analytical treatments that have been developed for mortality. It may be taken as a model for the development of a system of measures which can do much to enhance knowledge and understanding of family processes using data that are now available in the federal government.

The resources and techniques exist also for annual series of indexes of the level of residential segregation by race, and other topics. What is missing in these and other areas is a mandate and support for the development of analytical measures of major social trends comparable to the emphasis given economic trends.

Short of new analytical measures, much can be done to communicate the meaning of the masses of data already produced. Five years ago, Beverly Duncan showed how much more can be learned from educational attainment data by retabulating them in a cohort format.[16] There now exist powerful techniques of multivariate analysis which render obsolete much of the traditional inspection and oversimplified summarization (means, percentages, standardization on one variable out of many) which has characterized the analysis of complex cross-tabulations of social data.

In economic time series, our official agencies think nothing of adjusting the data to eliminate seasonal influences; they have done it for so long that their readers expect it, and the press relays reports of seasonally adjusted data routinely. The analogous operation for complex cross-tabulations of social data is readily available through multivariate analysis. We have recently received a Norwegian analysis of health data by the new techniques.[17] There is no reason why U.S. agencies should not follow suit. Failure to do so results, often enough, not merely in failure of communication, but in distortion of meaning.

Summary and Implications

We believe it is a mistake to identify the usefulness of social statistics for policy with the question of whether statistics meet the perceived information needs of federal policymakers. Such an identification probably assumes too great an ability by policymakers to identify their data needs. It clearly undervalues the contribution of statistics to the definition and anticipation of problems as distinguished from their solution. And it undervalues the role that other constituencies for statistics play in the policymaking process. We have urged, in particular, that statistical agencies apply accepted statistical techniques for the purpose of communicating far better than they now do, the meaning of the data they collect. And we have urged that the data be developed to a point where they more readily illuminate the problems of states and localities as well as the federal government.

These observations have several implications for the conduct of statistical work both inside and outside the federal government.

1. Analytical Function -- The communication function involves a good deal besides adequate staffing of public information offices, valuable as that may be. It is in part, and irreducibly, a statistical function, involving the preparation of analytical measures, the exploitation of models of social processes, and the compilation and analysis of time series from all competent sources to describe a topic. Such work requires a continuity

of attention that only an organization can provide, but it requires an organization which is committed to measurement and analysis as distinguished from information collection.

Both of us have served on the U.S. National Committee on Vital and Health Statistics. We recall the discussions from time to time in that committee, of the need for a "two-tiered" statistical agency, one part devoted to information collection, the other part to analysis and to communication in the sense in which we have used that term. The separation of function recognizes the fact that in an information-collection agency, the imperatives of information collection and data production invariably take priority over analysis whenever there is a choice, as there always is. The separation of function also recognizes what the Statistics Commission recognized--the disinclination of data-producing agencies to devote time and attention to exploiting the potential of data other than their own. The three compendia by Abbott Ferriss, on changes in education, the family, and the status of women, show what can be accomplished by statistical publications organized by topic rather than by agency.[18] The Statistical Abstract, OMB's forthcoming compendium on social indicators, and topical publications by the Census Bureau's Population Division cover the output of several agencies. But these publications do little or nothing beyond reporting data as reported by the originating agencies, they do little standardization, and they undertake no analytical measures. In general, Ferriss' lead has not been followed. The rule seems to be that the communications mission of data-producing agencies is defined by the data-collection mechanisms they operate. In the analytical tier of a two-tiered agency, the mission is not defined by a mechanism but by a set of topics or problems.

2. Involvement of University Research -- The analytical function need not necessarily be housed in a government agency, although there is much to be said for it, including access to the staff who designed the surveys, supervised processing, and designed record layouts and tabulations. Beyond this, government agencies have longevity, and the accumulation of statistical expertise, knowledge, lore, and professional tradition that goes with it. That kind of longevity is essential, but is rarely found outside government. The National Bureau of Economic Research has it, and economic statistics was the beneficiary of the years of tender loving care the Bureau devoted to the analysis of economic time series. With the uncertainties of project funding, few university-based research centers are in a position to provide the continuity of attention necessary to a similar endeavor in social statistics.

If that continues to be the case, there are nonetheless critical ways in which the involvement of university researchers in the social statistics enterprise needs to be cultivated. The analytical effort we have suggested presupposes that government agencies will actively draw upon the capabilities of university-based researchers to develop concepts, measures, and models of social processes. We have given a few illustrations of opportunities in this regard, and there are many more.

Closely related to this is the need to actively provide researchers with the materials they need for the analytical work from which developments in conceptualization, measurement, and modeling emerge. Some of this, of course, is provided by regular publications of the statistical agencies. But increasingly, scientific development presupposes access to basic data files, as illustrated by the work of Featherman and Hauser, of Allen, and of Winsborough, exploiting recent files of the Current Population Survey. The utilization of such files, while scientifically rewarding, can be extremely arduous, as users of the CPS files well know. Except for the public-use samples of the decennial censuses, it is generally the case that the micro-data files from government surveys are formatted and documented to the extent necessary to enable an in-house programmer who is experienced with the survey to use them, and no farther. This places tremendous burdens on the user, greatly restricts the number of users, and effectively impedes much of the analytical work which is needed to enhance the meaningfulness and communicability of the agencies' own data outputs.

3. Exploitation of Data Resources -- Much that we have said falls under the general rubric of exploiting existing data resources to improve knowledge about social change. In closing, we take this opportunity to return to a theme developed earlier. That is the utilization of data resources to assist in the statement and analysis of problems affecting areas below the national level. We have elaborate models of economic processes at the national level. Rudimentary beginnings have been made on a few models of social processes at the national level; that work needs to be encouraged and continued, and the validated results of it ought to be sought out and utilized in government social statistics programs.

At the state and local level, where a great deal of social policy is made, we know of few examples of subnational model building which lends significance to local data.[19] Model building for any one local area is beset by the difficulty that any area is related to others in a multiplicity of systems. Hence it is necessary to study systems of areas. A suitable place to begin would be with data capable of describing, for a set of communities, changes in a variety of facets of community life--employment, industrial composition, population and vital statistics, education, crime, and so forth, and to begin to seek structure in the way the data array themselves and in the relationships of change on one characteristic to change on others.

Over a period of time, such work might yield relationships which would lend far greater meaning to local observations than is possible at present, for it would permit the explanation of local phenomena by reference to the behavior of the systems of which a community is a part.

Figures on the migration patterns of a community, its educational level, or conceivably its crime rate, could be approached not only by asking "What was it last year" or "What is the measure for similar communities" but by asking "What did you expect?" in terms of the measure's relationships in a system of measures.

Why should not the tape files of the County-City Data Book be exploited for such a purpose? In these files we have quinquennial readings since 1952 for counties, cities, and metropolitan areas, on a wide range of key characteristics, which ought to be applicable to the problem we have posed. To be sure, a good deal of work would be required to ensure consistency of items from one year to the next, to document inconsistencies, and for other preparatory tasks. But the fruit of such efforts should be substantial. If this approach is judged not optimal, others have been proposed for consideration. In any case, it is the task which interests us at the moment, rather than the means for its accomplishment. The task is the development of measures and of models that lend them meaning, and for this purpose to exploit the resources at hand. Here, the role of the statistical agencies is to facilitate the exploitation of the data, and to incorporate in their own programs the validated results of the analysis such an effort makes possible. The policy contribution is to assist in the understanding of social processes and the definition and anticipation of social problems.

Footnotes

1. President's Commission on Federal Statistics, Federal Statistics, Vol. I, Washington, D.C., 1971, Chapter 3.

2. Ibid, p. 89, quoted in Otis Dudley Duncan, "Federal Statistics --Nonfederal Statisticians," American Statistical Association, Proceedings of the Social Statistics Section, 1972.

3. Ibid, p. 82.

4. Albert D. Biderman, "Information, Intelligence, Enlightened Public Policy: Functions and Organization of Societal Feedback," Policy Sciences 1, 1970, p. 217-230.

5. Peter J. Henriot, "Public Policy Implications of Social Indicators Research and Development," unpublished paper presented at Symposium on Social Indicator Research in an Urban Context (Ottawa: Ministry of State for Urban Affairs, 1973).

6. Robert L. Berg, ed., Health Status Indexes, Chicago, Hospital Research and Education Trust, 1973.

7. David L. Featherman and Robert M. Hauser, "Design for a Replicated Study of Social Mobility in the United States" in Social Indicator Models, edited by Kenneth Land and Seymour Spilerman, Russell Sage Foundation, forthcoming.

8. Halliman H. Winsborough, "Age, Period, Cohort, and Education Effects on Earnings by Race -- An Experiment with a Sequence of Cross-Sectional Surveys" in Social Indicator Models, edited by Kenneth Land and Seymour Spilerman, Russell Sage Foundation, forthcoming.

9. W. M. Mason and R. W. Hodge, A Proposal to Use the Tapes from the March Current Population Surveys to Study the Changing Socioeconomic Location of Blacks and Women in the United States (draft, 1973).

10. National Education Association, NEA Research Bulletin, 1971, Vol. 49, No. 3.

11. William Alonzo, "The System of Intermetropolitan Migration Flows," in U.S. Commission on Population Growth, Research Papers, Vol. 5, Population Distribution and Policy, Sara Mills Mazie, ed., Washington, D.C: G.P.O., 1973.

12. U.S. Commission on Population Growth and the American Future, Final Report, Population and the American Future, Chapter 3, Washington, D.C.: G.P.O.,1972

13. President's National Goals Research Staff, Toward Balanced Growth: Quantity with Quality, Washington, D.C.: G.P.O., 1970, p. 60.

14. Abbott Ferriss, Indicators of Change in the American Family, New York: Russell Sage Foundation, 1970.

15. Kenneth C. Land, "Some Exhaustible Poisson Process Models of Divorce by Marriage Cohort," Journal of Mathematical Sociology, Vol. 7, 1971, p. 213.

16. Beverly Duncan, "Trends in Output and Distribution of Schooling" in Eleanor Bernert Sheldon and Wilbert E. Moore, eds., Indicators of Social Change: Concepts and Measurements, New York: Russell Sage Foundation, 1968, Chapter 8, pp. 601-672.

17. Asmund Langsether, "Indicatorer for Helsetilstanden I Befolkingen: En videre analyse av Statistisk Sentralbyras helsundersøkelse. INAS Report No. 71-3 of the Institute of Applied Social Research. Oslo, February, 1972.

18. Abbott L. Ferriss, Indicators of Trends in American Education, 1969; Indicators of Change in the American Family, 1970; Indicators of Trends in the Status of American Women, 1971. New York: Russell Sage Foundation.

19. James G. Anderson, "Causal Models and Social Indicators: Toward the Development of Social Systems Models," American Sociological Review, 33 (1973), p. 285.

DISCUSSION

James McPartland, The Johns Hopkins University

The Parke and Sheldon paper raises many important points concerning the relationship of social statistics and public policy. I would like to give emphasis to some of their ideas by applying them to a different example: the state and local educational accountability and assessment reports which have developed in the past couple of years and seem to promise to become quite widespread in the next several years.

State Educational Assessment Reports

A recent survey by ETS finds that 30 states are now operating educational assessment programs.[1] Sixteen are required by state educational accountability laws to do so. In the other states, the idea was introduced by the statisticians and professionals in the education agencies. These assessment programs and reports use standardized achievement tests and exercises for students, sometimes tests constructed by the states themselves.

Generally, the reports of the test results are presented in one of two ways: either (a) unadjusted school or district mean achievement scores, or (b) adjusted residual mean achievement scores (adjusted for student inputs via measures of student socio-economic status, previous year's test scores, or some so-called ability test).

Applying Four Ideas from the P-S Paper to this Example

There are four points from the Parke-Sheldon paper which I will apply to this example.

First, and most obvious, this example shows the need to "value the importance of policy problems of states and localities" as stated by Parke and Sheldon. These local laws and activities have come about because of the clear need for better information to justify and allocate the local monies which go to public education. (This expenditure is now more than 40 percent of total state and local outlays, on the average.)

However, this need is not likely to be solved in this case simply "through projections of (federally collected) social data for areas below the national level," but seems to require the actual development at the local level of separate statistical collection and reporting activities. In other words, the example of educational accountability emphasizes the need expressed in Parke and Sheldon, but also suggests that not all of the "action" is at the federal level, with federal statistical results projected and reported at the local level. In this example, while national programs on school tests are helpful (such as National Assessment of Educational Progress or the development of the "anchor test", which are partially supported by federal agencies), it seems for political and practical reasons that the local agencies are needed to do the real work of developing statistical indicators and report formats. At any rate, much of the leadership in this development has come from statisticians working on the state and local level.

A second need expressed by Parke and Sheldon --to value "the role of public understanding in policy development"--and their prescription for statisticians to develop indicators and reports accordingly, is of great importance in the state assessment example.

Each of the two approaches to state educational assessment I mentioned (unadjusted school means and means adjusted for student inputs to schools) are usually faced with erroneous public interpretations unless special precautions are taken in reports.

The first kind of statistic (the unadjusted school means) tells something about the general level of learning in the school-aged population; and a comparison of schools on this measure locates where the poor students go to school. But this unadjusted statistic does <u>not</u> say anything about program evaluation, i.e. which schools are doing the best job and why. This is because the average student achievement scores in a school have been found to be more a function of family background and previous experiences of the students, rather than a function of the school's program itself.

As obvious as this may be to this audience, a large fraction of the public (and legislators) misinterpret the unadjusted scores as saying more about the quality of various school programs than about where the poor students happen to go to school. So, statisticians working or advising on the state reports need to be careful to prevent this misinterpretation, perhaps by presenting unadjusted results only in aggregation above the school level. In doing so, the reports educate the public that schools are only one of several influences on children's learning.

The second kind of statistic (adjusted residual school means standardized for student inputs) must face another set of public misunderstandings. One difficulty arises from popular assumptions about the particular input variables which should be controlled. A second problem comes from the degree of precision which may be assumed for the results.

Testing specialists generally agree that most of the so-called ability, aptitude or IQ tests which are administered to large groups are simply <u>other</u> achievement tests (that is, there is no basis on which to choose one test as more indicative of native aptitude than another, even though one test may carry the label "aptitude" or "ability" test). This means that one should <u>not</u> use a concurrent "ability" test as an input control in analyzing residual achievement influences of school programs.[2] Nevertheless, the

essential similarities of many "ability" and "achievement" tests are not well known by the public or by many educational professionals. Unless statisticians and measurement specialists assert their proper role in developing report and analysis methods, essentially technical questions will be decided politically--by votes of lay advisory committees--and important public misunderstandings will go uncorrected.

In addition, the public tends to make a great deal over small differences between schools in adjusted scores, even though many of the differences may be well within the range of likely random error. As suggested by a recent RAND report on educational outcome measures, this kind of public overinterpretation can be minimized by using a coding scheme of broad categories to present results, rather than suggesting a psuedo-exactness by employing continuous cardinal residual measures.[3] In RAND's words, "Crude measures should be employed crudely."

Finally, the example of state activity in school assessment serves to emphasize two other ideas offered by Parke and Sheldon. Now, I refer to their comments on "measures which derive their significance from models of social processes" and also their remarks concerning reports and indicators which narrow the range of options rather than point definitively to a single policy choice.

Underlying the approach using adjusted means is a general model of student input - school learning processes - and learning outputs, but in a simplified version. A major goal in using the adjusted means for a comparison of school programs is to learn which aspects of the total instructional program work best for specific learning outcomes: for example, is reading most affected by the individualization of assignments, by the kinds of staffing, by the way grades and rewards are managed, by the grouping of students? In general, what elements of the school's program are most important for specific outcomes, and thus require more emphasis and future investment in the least successful schools?

To truly answer this question analytically requires a complicated causal model that includes all student input factors, plus all the possible school factors that may distinguish one school program from another, expressed for each separate kind of learning. Clearly, such a specific model is not likely at present: it is beyond our present knowledge to specify all the variables, and the various factors are probably not statistically distinguishable across schools but are highly correlated with one another. Even if we knew the variables, the data requirements to implement such a model are beyond the capacities of most state and local agencies to collect and analyze.

So, in those state plans that include residual adjustments for student input, a more simplified model is used to initially narrow the range of possibilities to explain present differences in school effectiveness. This simplified model seeks only to identify exemplary schools by name, but not (at first) to specify the elements of their programs which make them exemplary. (The simplified model is to regress achievement on student background factors, and then identify schools which are consistently above or below the expected level on several indicators.) Thus, a general social model is in mind in developing the approach, but to begin with, a simplified model is used and the information gathering process is divided into separate stages.

After the exemplary schools have been identified, data can be collected formally or informally about how their school programs and staff are actually different from the rest. This second step provides some direction to program changes which might be attempted and evaluated in the least successful schools, and incorporated in more sophisticated models of school effects for later reports.

Thus, both of the points raised in the Parke-Sheldon paper (using social models, and the virtue of narrowing the range of possibilities through practical statistical indicators) seem to be a major part of some state and local educational assessment programs.

My comments on the Parke and Sheldon paper reflect that many of their main points on the relationship between social statistics and public policy seem to apply very well to some important recent state and local policy concerns and statistical projects in the field of education. The example I chose also suggests that federal statistical agencies may not be the only locations where important developments are underway: state agencies and localities also are at work on the problems of social statistics and public policy.

Footnotes:

1. State Educational Assessment Programs, 1973 Revision. Princeton: Educational Testing Service, 1973.

2. Dyer, Henry S., "Toward objective criteria of professional accountability in the schools of New York City," Phi Delta Kappan, 52, 1970, 206-211.

3. Klitgaard, Robert E., Achievement Scores and Educational Objectives. Santa Monica: Rand Corporation, 1973.

THE WORK OF THE COMMITTEE ON NATIONAL STATISTICS

Margaret E. Martin, Committee on National Statistics

The Committee on National Statistics was established by the National Research Council (the operating arm of the National Academy of Sciences) in response to growing concern about the adequacy, validity, timeliness, and utility of statistical procedures and information central to major national decisions.

Although the NRC has had a long-standing interest in the quality of the statistical information basic to public policy formation, the immediate impetus to establish the Committee came from a major recommendation of the President's Commission on Federal Statistics, that an independent, continuing group be established to review the federal statistical system and recommend improvements. The Commission reported:

"We are convinced that . . . , a need exists for continuous review of federal statistical activities, on a selective basis, by a group of broadly representative professionals without direct relationships with the federal government."

and urged:

" . . . that an NAS-NRC committee be established to provide an outside review of federal statistical activities." 1/

The Committee, first appointed in January, 1972, is attached to the Division of Mathematical Sciences in the National Research Council. Members are appointed for three-year terms and serve without recompense. Current members of the Committee are William Kruskal, chairman, Douglas Chapman, Morris Hansen, Stanley Lebergott, Frederick Mosteller, I. Richard Savage, Elizabeth Scott, William Shaw, and Conrad Taeuber. In addition, Cuthbert Daniel and Bernard Greenberg served through the spring of 1973. The Committee now has a professional staff of three, all serving part-time, Margaret Martin, Hyman Kaitz, and Edward Tufte.

The Committee and its staff represent a broad spectrum of disciplines in which statistics are applied, as well as a variety of experience in statistical methodology, data collection, and data analysis. As a committee sponsored by the National Academy of Sciences, it is primarily oriented toward giving scientific advice to the federal government. Yet even its name indicates a somewhat broader interest — "national" statistics rather than "federal" statistics is intended to connote any statistics of important public concern, whether collected by the federal government or not.

Coming to the Committee months after it had been established, I searched for a detailed statement of its functions or a "frame of reference" to help guide future planning. The most direct statement was in the President's Commission's report and might be summarized in three words as, "Carry on, chaps."

In these circumstances one might think that the early attention of the Committee would have concentrated on building such a framework and establishing priorities within it. However, it early became apparent, with the wide range of backgrounds and experience represented on the Committee, its infrequent meetings, and the multitude of urgent statistical problems pressing for attention, that we might better proceed with a few critical projects immediately. We were in danger of bogging down in disucssions of generalities.

Partly, the functions of the Committee on National Statistics may be viewed in terms of what the Committee is not. It is not a group representing the profession, as does the American Statistical Association; it is not a group representing users, as does the Federal Statistics Users' Conference; it is not intended to duplicate the functions of the federal statistical agencies, nor of academic research, nor of commercial or non-profit contracting organizations. Rather, it is expected, by selecting significant, broadly applicable projects and approaching them in a creative, multidisciplinary fashion, to focus the expertise of specialists outside the government on important statistical issues. At the same time, these projects must be such that most of them can be funded by the specific agencies which will be the recipients of the advice; and the process of developing the projects must be reasonably compatible with the style of operation of the National Research Council.

Despite these rather formidable constraints, there is no lack of suitable issues to begin on. We are constrained, rather, by the smallness of our staff, the limits of our expertise and the interests of our members and possible sponsors. The National Research Council has given us a comfortable home and all manner of supportive services; the Russell Sage Foundation has given us an initial start-up grant to get going and show what we can do. So where are we?

We are in that suspenseful period between project conception and initial approval on the one hand and the actual transfer of funds on the other. We have two projects now in the final stages of

consideration by the sponsors. Several more are in various stages of preparation.

Once we have an approved project, we plan to appoint a panel to carry forward the study and to hire appropriate supporting staff. The end product would be a report of findings and recommendations suitable for public dissemination.

We do have some general funds, and hope to continue finding resources to support some small activity at the Committee's own option. Our goal is to reach a "steady state" in which we might be working simultaneously on four projects in various stages of completion. Of the four, we would hope that three would be funded specifically by federal agencies or other sponsors; the fourth would be undertaken by the Committee — perhaps something so broadly applicable no one agency would be a suitable source of funds; perhaps a small-scale exploratory investigation which might develop later into a separately funded project.

So far I have sketched organization, functions and modus operandi in general terms. Specific project plans are of more interest and that is what I shall spend the rest of my time describing. I must preface all of these plans, however, by noting they are dependent on final actions by others. We have rushed at high speed more than once during the past year to submit proposals to meet the requirements of potential sponsors, only to find that for one reason or another action on the proposal has been delayed. It reminds one of that description of army life — "Hurry up and wait."

I should like to describe briefly the four projects that are farthest along. The first should be of particular interest to the Social Statistics Section. Five times in little more than a decade the Section has sponsored a program at the annual meetings on the statistical needs in the law enforcement and criminal justice area. 2/ This interest has been matched elsewhere with the result that the Law Enforcement Assistance Administration has been given the responsibility for making major improvements. The most innovative statistical project is the National Crime Panel, sometimes referred to as victimization surveys. These are surveys of samples of households and of businesses to obtain reports of crime from the victims.

The LEAA, with the assistance of the Census Bureau, has undertaken a program of methodological testing, following the pathbreaking efforts undertaken for the President's Crime Commission in 1967. The preliminary testing and experimenting phase was completed more than a year ago, and, with major decisions made, a national sample of households was drawn and enumeration started in mid-1972. Enumeration of a sample of commercial establishments commenced a few months later.

We have a project proposal, designed at LEAA's request, to evaluate this National Crime Panel, not only from the point of view of its statistical methodology, but also of its utility. How does such a statistical undertaking serve the variety of users, and in particular, the needs of social scientists? Our proposal, which is awaiting final action at LEAA now, envisions a cooperative review by the Committee on National Statistics and the Academy's newly organized Assembly of Behavioral and Social Sciences. We shall also keep in close touch with the Social Science Research Council's subcommittee on Criminal Statistics.

Once the go-ahead signal is given, the Committee will appoint a subcommittee or work group to engage in the actual review, with at least one member of the parent Committee participating, together with specialists in the sociology of crime, statisticians, and other experts. One member of the Committee on National Statistics, Conrad Taeuber, has already agreed to act as chairperson of the subcommittee.

The National Crime Panel raises many challenging methodological questions, questions related to memory bias and recall problems, ability of one respondent to report for another and the specifics of question-wording on sensitive and easily misunderstood subjects. Many of these questions have been tested in preliminary surveys by the LEAA and the Census Bureau prior to making major decisions for the national panel. Thus the present survey of households uses a six-month recall period, because it was found too much was forgotten when the period was lengthened to a year; it interviews each respondent for himself rather than relying on a single household respondent to report for all members; and it uses the technique of "bounded" interviews to prevent "telescoping"— that is, to overcome the tendency of respondents to report major events as having fallen within a specified period when they actually occurred earlier. This telescoping effect can be overcome if a first interview is made, events recorded and then six months later on, during a second visit, reported events are edited on the spot by the interviewer against events reported earlier. The second interview is thus "bounded" by the first, and statistics are compiled only from the second and succeeding interviews. Thus the decisions to use 6 months rather than 12 as the recall period, to insist on interviewing respondents each for himself, and to "bound" the interviews have all served to increase both the quality and the cost of the household survey. For questions such as these, on which considerable methodological work has already been done, I anticipate that the evaluation group might

simply review the evidence and assess the decisions.

Many other aspects may have been less thoroughly considered. This survey, which will provide quarterly estimates of the incidence of crime and much more detail by type and characteristics on an annual basis, will form the basis for time series of numbers and rates of events which inevitably will be compared with data from administrative systems — crimes reported to the police. Specialists already know that there are significant differences in level and one may anticipate at least occasional differences in rates of change from period to period. A review of how best to present and interpret such apparently conflicting evidence to various groups of users might be pursued by the subcommittee. Many other aspects of data analysis will no doubt be considered.

Less fully explored and possibly less obvious are the difficulties of obtaining descriptions of events so precisely that uniform legal concepts can be applied as consistently as possible across public jurisdictions, among different social groups and through time. The National Crime Panel depends on the perception as well as the memory of respondents that certain types of events have occurred. The evaluation group may want to look closely at such issues. Respondent ignorance affects the reports from business concerns in major ways. At the present time, for example, the business reports do not cover types of crimes which are unlikely to be discovered and reported currently as discrete events. Thus robbery is included but shoplifting and employee theft are not, since they remain largely unknown until they are the presumed reasons for "inventory shrinkage".

In addition to evaluating such decisions and exploring alternatives, the evaluation group will be asked to stretch its multi-disciplinary horizons and consider the utility of the entire undertaking. Who will use the results? How? Would some changes in directions or emphasis be worthwhile? One of the most exciting aspects of this project to me is that this is the first occasion of which I am aware in which an independent outside group has been asked to evaluate a major new statistical system at a very early stage — so early, in fact, that no data have yet been published from the national panel.

I mentioned earlier that the Committee on National Statistics hopes to sponsor one project of its own in addition to those funded specifically by others. One such project has been approved and work on it will shortly get underway. This is a project proposed by our consultant, Hyman Kaitz, in which he would deal with several aspects of the problem of using data, subject to errors of various kinds, in the preparation of press releases. He will consider these problems in connection with economic time series and would deal initially most particularly with the monthly employment and unemployment releases, as an example.

Kaitz is thinking of both the statistical and analytical considerations here and that sometimes they appear to conflict. Geoffrey Moore described some of these conflicts in a recent article. 3/ One example was the difficulty of describing to the general public an increase in the white unemployment rate of, say, three-tenths of a point, because it is statistically significant, yet not mentioning a statistically insignificant increase of seven-tenths in the black rate. Kaitz plans to explore this and a number of related questions under 5 headings:

(1) A review of past practices against the consistent application of known criteria;

(2) A review of the significance criteria now in use and an examination of alternatives;

(3) Research on the impact of seasonal adjustment procedures on criteria of significance;

(4) Research on alternatives to the measure MCD, months for cyclical dominance, as a tool in business cycle analysis;

(5) Research on how statistical and other significance criteria are translated into ordinary prose for the public and how this process of communication might be improved.

He plans to prepare separate reports on these five sub-topics for circulation and comment by time series analysts. Some of the papers may be exploratory. It is possible that some might develop at a later stage into full-scale projects for which outside funding might be sought. In any case it is expected that wider dissemination of his results will be made following expert review.

Still another project draws in an entirely different area of application. Another committee at the Academy, the Climatic Impact Committee, has asked our assistance in interpreting the fragmentary and conflicting evidence on the incidence of skin cancer. They wish to establish what the relationship between skin cancer and latitude is, as part of an analysis of the possible impact that widespread development of SST planes might have, a question on which they are advising the Department of Transportation. The chain of reasoning seems to be — operation of SST planes may deplete ozone in the upper atmosphere causing an increase

in ultraviolet radiation and skin cancer. Since the average amount of upper atmosphere ozone decreases as latitude decreases, the effect of ozone depletion on skin cancer may be calculated if it is known how the incidence of skin cancer varies with latitude.

The direct effect of latitude is confounded with many other variables — some natural phenomena such as altitude or smog — some demographic such as migration, race, or ethnic group (blacks seldom if ever suffer skin cancer; Scandinavians are not nearly as susceptible as those of Celtic origin) — some social effects, amount of clothing, sun bathing practices — and economic effects, particularly occupations requiring considerable outdoors work as in farming or maritime occupations. Problems of identifying and enumerating cases of skin cancer also present unusual difficulties.

The Committee on National Statistics will attempt to resolve the differences between two existing data sources for parts of the U. S., will comment on the adequacy of the data for correlating latitude and skin cancer incidence, and possibly will make recommendations for developing more adequate information in the future. Present plans are for a quick investigation and early report. Depending on the nature of the findings, the Committee may recommend further work on developing an improved data system.

When the President's Commission on Federal Statistics recommended that an independent advisory committee be set up under the aegis of the National Academy of Sciences, it said in part, "Such a body could monitor the implementation of Commission recommendations and, even more important, conduct special studies on statistical questions it deemed important because their favorable resolution would contribute to the continuing effectiveness of the federal system." 4/ One of the tasks of the Committee and its staff has been to fit this general prescription into the pattern of operation of the NAS-NRC, which relies mainly for support on funding for specific projects by specific federal agencies.

In part, we have attempted a solution by looking for partial support from other sources, from private foundations, and here we have found both practical support and warm encouragement from the Russell Sage Foundation. But beyond this, I believe that we shall find that we can develop a series of specific projects suitable for agency support which will deal with various aspects of a more general problem. For example, we have already developed a proposal which HEW is considering that would examine what statistical methods might be used to improve the process of determining user needs for statistics and establishing priorities among them. This would be an exploratory survey, using one statistical center as a case study. Whether or not that specific project is approved, I am sure that the Committee will be dealing with the central issue from various standpoints as time goes on. Already, we have made a beginning at our last meeting of the Committee in Washington by bringing together representatives of economic policy agencies and those responsible for economic statistics for an informal discussion of how priorities are set for economic statistics. Even this brief introduction to the problem evoked a number of suggestions and comments which the Committee no doubt will wish to examine in the future — for example, a proposal that the Statistical Policy Division of the Office of Management and Budget should have some responsibility for developing estimates and projections of important economic variables so that it would be more sensitive to the most important statistical gaps when exercising its statistical planning functions; a suggestion that the "mission-oriented" agencies should be given funds so that they could buy the statistics they need from the statistics bureaus without interfering with the on-going general-purpose statistical series; and finally, reiteration of the basic conflict, in terms of budget and manpower resources, between the needs for national statistics important for policy purposes and the needs for detailed information for local administration of many public programs. This last point brings us full circle — I am sure that the demand for detailed local-area statistics would be one of the most important questions to be explored in the HEW project, should it be funded. Whether or not the Committee gets into some of these issues and their possible solutions in that instance, I am sure it is an area in which, sooner or later, the Committee will do some work.

Another general problem of interest to the Committee is that of confidentiality of statistics. The Committee has not yet determined on any plan of inquiry, but preliminary discussions are being held with the Bureau of the Census to see if we can outline a project of mutual interest.

In summary, we find no lack of possible topics on which the Committee might make a contribution. Our problem is rather one of making an intelligent selection of the most important — of those which are uniquely suitable for a Committee such as ours; which are feasible; in which we believe we can make a real contribution; and last, but by no means least, those with a likelihood that the recommendations of the Committee will be given serious consideration.

In closing, I should like to make just one more observation and perhaps correct a misapprehension arising from my earlier remarks. I said that when I started as executive director for the

Committee staff I found no blueprint of the Committee's functions. After nearly a year, I have produced no blueprint myself, yet looking back at the talk our chairman, William Kruskal, gave more than a year ago to a Federal Statistics Users' Conference, and since reprinted in Science, 5/ I find most of what I wished to say already said — more succinctly, more elegantly and more imaginatively. The Committee is indeed fortunate to have such a perceptive, hard-working and dedicated chairman. Without that kind of leadership, it is hard to see how we could succeed.

REFERENCES

1/ Federal Statistics, Report of the President's Commission, Washington: U. S. Government Printing Office, 1971, Vol. I, p. 175, p. 4.

2/ Sessions at the Annual Meetings of the American Statistical Association reported in Proceedings of the Social Statistics Section for the appropriate year:

"Statistics of Crime and Delinquency: Progress in Measurement", 1963, pp. 1-18.

"Statistics in Criminology", 1965, pp. 1-35.

"Improving Federal Statistics on Crime and Criminals", 1968, pp. 101-116.

"Developments Toward a National Criminal Statistics System", 1971, pp. 106-125.

"Victimization Surveys as a Source of Crime Statistics", 1973, elsewhere in this volume.

3/ Geoffrey Moore, "On the 'Statistical Significance' of Changes in Employment and Unemployment", Statistical Reporter, March 1973, No. 73-9, pp. 137-139.

4/ Federal Statistics, Vol. 1, p. 175.

5/ William Kruskal, "The Committee on National Statistics", Science, 22 June 1973, Vol. 180, pp. 1256-1258.

APPLICATION OF MULTIVARIATE REGRESSION TO STUDIES OF SALARY DIFFERENCES BETWEEN MEN AND WOMEN FACULTY

M. G. Darland, S. M. Dawkins, J. L. Lovasich,
E. L. Scott, M. E. Sherman, and J. L. Whipple

University of California, Berkeley

SUMMARY

Women who are employed receive lower salaries, on the average, than men. We investigate to what extent the differences in faculty salaries can be explained by relatively objective factors, such as lack of the Ph.D. and differences in performance, and to what extent they appear to be the result of discrimination. Using regression on more than 25 predictor variables, we estimate the salary of all faculty, of men faculty, and of women faculty in various types of universities and colleges and in various fields.

We find that the predictors that are important in determining salary are not always the same for men and women, but that both salaries are well predicted with a typical R of 0.8. When we compare the estimated salary of a man and a woman of the same abilities and performance, or when we compare the salary a woman actually gets with that predicted from the men's equation, we find that women tend to be underpaid by about $1500 annually, on the average, and often by much more. The amount of underpayment is more pronounced in the research universities, in the biological and physical sciences, and at the higher levels--just where salaries tend to be highest and women are scarce.

The apparent discrimination in faculty salary due to sex is strong and persists for every race. The Carnegie survey data indicate that the salary differential due to sex is much larger than that due to race.

1. INTRODUCTION

Women are underrepresented on the faculties of colleges and universities, and those women who are employed tend to be paid less well than the men at the same type of institution. The figures published by the Office of Education and by the National Education Association [1] show that the median salaries are between $1,000 and $4,000 lower for women than for men in similar institutions. The differences in median salaries tend to be larger in the more selective colleges and universities, just those where the salaries themselves tend to be higher. However, one must consider that fewer women have the doctorate, that women who are employed in academe have a different age distribution than men, and that they tend to be concentrated in certain fields. In this paper we estimate how much of the observed salary differentials can be explained by relatively objective factors such as highest degree held, differences in performance and in attributes, and to what extent they appear to be the result of sex discrimination. Do men and women of the same ability and performance receive different salaries, and if so, by how much?

The data available are from the large-scale national survey [2] made in 1969 by the Carnegie Commission on Higher Education in cooperation with the Office of Research of the American Council on Education. This survey included a comprehensive questionnaire returned by 60,028 faculty members located in 78 universities, 168 four-year colleges, and 57 two-year colleges. Astin and Bayer [3] used a linear regression equation with 32 predictor variables to compare the average salary a man would receive with the average for a woman having the same rank, background, and achievements. They included indicators of the type of institution and field among their predictor variables; they found that the average discrimination exceeds $1,000 in salary and one-fifth step in rank.

In an effort to get at the source of discrimination in more detail and also to obtain better estimators, we extended [4] the study of Astin and Bayer to investigate the salary differences of men and women for different types of institutions (as set out by the Carnegie Commission [4]) and different fields separately. We also included higher order interaction terms and an indicator of full-time versus part-time employment. Our analyses were based on the replies of all women sampled and a 25% random sample of the men. We used the salary intervals of varying width (averaging $3,000) adopted by the Carnegie

Table 1. COEFFICIENTS OF THE MULTILINEAR REGRESSION EQUATION FOR PREDICTING FACULTY SALARIES (IN $1,000)

Research Universities I - Biological and Physical Sciences - 1183 Men, 312 Women

	Men and Women	Men	Women	Variable
Const.	10.19	9.13	3.11	
1	-1.48**	---	---	Sex, 1 = male, 2 = female
2	.49***	.57**	.44*	Date of birth, 1 = 1908 or before to 9 = 1944 or later
3	.96*	.75	-.52	Marital status, 1 = never married, 2 = married or formerly married
4	1.16***	.74**	1.04*	No. of children, 1 = none to 4 = three or more
5	2.01***	2.23***	1.98***	Highest degree, 0 = BA or less, 1 = MA, 2 = doctorate
6	-.52***	-.59***	-.30*	Year of highest degree, 11 = 1928 or before to 21 = 1967 or later
7	.28	.24	.11	BA from a prestigious school, 0 = no, 1 = yes
8	-.31	-.24	-.86**	Graduate degree from a prestigious school, 0 = no, 1 = yes
9	.25	.31	.01	Support toward highest degree, 0 = none to 2 = TA/RA plus fellowship
10	---	---	---	Rank (variable omitted)
11	.75***	.86***	.46*	Years employed in academe, 1 = one or less to 8 = 30 or more
12	-.53***	-.59***	-.30*	Yrs empl'd in present inst'n, 1 = one or less to 8 = 30 or more
13	-.22⁻	-.30⁻	.07	Quality of present inst'n, 1 = high to 7 = low
14	1.29***	1.28***	1.25***	No. of articles, 1 = none to 6 = more than 20
15	.44***	.35*	.51⁻	No. of books, 1 = none to 4 = five or more
16	.19	.25	-.07	Assoc'n with a research institute, 1 = yes, 2 = no
17	.06	.01	.51*	No. of sources of research support, 0 to 6
18	.90***	.84***	1.07**	No. of sources of paid consulting, 0 to 6
19	-.31*	-.43**	.02	Research/teaching inclination, 1 = heavily research to 4 = heavily teaching
20	.63***	.68***	.30*	Administrative activity, 1 = none to 7 = 81 to 100 percent time
21	.00	-.03	-.04	Consulting, 1 = none to 7 = 81 to 100 percent time
22	-.28*	-.27⁻	-.45	Outside professional practice, 1 = none to 7 = 81 to 100 percent time
23	-.44***	-.51***	-.02	Hours taught per week, 1 = none to 9 = 21 or more
24	2.40***	2.48***	2.24***	Salary base, 1 = 9/10 months, 2 = 11/12 months
27	-.15***	-.15***	-.17***	Interaction: date of birth and number of articles
28	-.22	---	---	Interaction: sex and number of children
29	-.13**	-.11*	-.18*	Interaction: date of birth and number of children
30	.00	.31	.26	Interaction: sex, marital status, and age, 1 = male, never married, under 30 to 8 = female, married or formerly married, 30 years or older
31	-2.06***	-2.07***	-.96*	Part-time by Rule 5, 1 = full-time, 2 = part-time

⁻ Individual coefficient differs from zero at 0.10 significance (two-sided); * at 0.05; ** at 0.01; *** at 0.001.

Survey. The present study differs from the last only in that the salary intervals were converted to dollars before any estimates were computed or comparisons made. Since we want the results in dollars, carrying out the computations in dollars avoids additional bias.

2. ESTIMATES OF SALARY

We find good estimates of faculty salary from a simple additive equation. The variables used are listed in Table 1 with three sets of coefficients estimated for a prediction equation for salary in thousands of dollars. The first set of coefficients shown corresponds to 28 predictor variables, of which 4 are interaction terms, and was estimated using the combined sample of men and women in the field of Biological and Physical Sciences of institutional type Research Universities I. The next set, obtained using only the men faculty, and the last set, using only the women, omit the 2 predictor variables that involve sex in such a way that they become redundant, thus retaining 26 predictor variables of which 3 are interaction terms. (Even though the last interaction term involves sex, it is not redundant and was retained as a measure of mobility.) We have used the same set of predictor variables for each combination of field, type, and sex. Since rank is tightly locked to salary in many institutions, it was not used as a predictor variable. Preliminary studies suggested that including rank would cloud the effects of other variables more suitably regarded as predictors.

Initially, stepwise regression was employed to aid in selecting predictor variables, and tests of linear hypotheses were performed for particular combinations of field and type so as to see what might be important in setting up the systematic analyses. For these initial investigations, versions of the programs LINWOOD [5] and BMD [6] were used. The main analyses were carried out with an adaptation of DANIEL, a local version of the LINWOOD program.

The asterisks following the estimated coefficients in Table 1, and in Table 2 below, indicate the predictor variables that are individually significant in determining faculty salaries in the specified field and type of institution. Since many of the predictor variables are presumably themselves collinear, the individual significance probabilities do not necessarily give a full picture of the importance of particular predictors. Nevertheless, they are of interest.

We note that when men and women are considered together, the variable sex is important and its coefficient is appreciable, - $1,480. Thus, when all other predictor variables are fixed, the predicted annual salaries for women are $1,480 less than those for men. However, because the predicting equations fit to the men and to the women separately differ very significantly, this joint equation should be discarded.

Examining the other two columns of coefficients, the set for men and that for women, we see that having higher degrees is very important for both sexes. Also important for both sexes is the number of articles published, the salary base, and the interaction term date of birth by number of articles. On the other hand, for some variables, the coefficients estimated for men are quite different from those estimated for women. The increase in salary with the period employed in academe is twice as much for men as for women (predictor No. 11), both men and women gain by changing institutions but men gain twice as much as women (No. 12), men also gain twice as much as women by administrative activity (No. 20), men who teach less are paid more but a woman's salary is unaffected by hours of teaching (No. 23), and men lose more than women by being employed part time (No. 31). (This last conclusion is uncertain because the survey did not ask whether employment was full time; we estimated [4] this from answers to other questions since it must be an influential variable in predicting salaries.)

The sets of coefficients for other fields and other types of institutions are shown in Table 2. On each page of the table we show the coefficients for various fields in an institutional type, starting with Research Universities I. There is considerable variation in the coefficients and loss of power with small samples. The general pattern persists with some exceptions: the number of books published (predictor variable No. 15) tends to be important in the field Humanities, especially for men, and books are more rewarding for men. Having children (No. 4) tends to decrease the salary of women if they are not in the sciences or when they

Table 2. COEFFICIENTS OF THE MULTILINEAR REGRESSION EQUATION FOR PREDICTING FACULTY SALARIES (IN $1,000) IN RESEARCH UNIVERSITIES I

Field:	Bio/Phys Sci.		Education		Fine Arts		Humanities		Soc. Sci.		New Professions	
Predictors Sex:	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Constant	9.13	3.11	9.49	16.47	24.49	11.26	11.36	12.89	16.98	9.44	18.99	9.57
2 Date of Birth	.57**	.44*	.03	-.12	-.04	-.11	.19	-.19*	.15	-.14	.04	-.05
3 Marital Status	.75	-.52	.31	-.37	.85	.25	.19	-.20	.38	-.72	.76	-.44
4 No. of Child.	.74**	1.04*	.61	.26	.88*	-.52	.60*	-.71*	1.04***	-.48	.49	-.26
5 Highest Degree	2.23***	1.98***	2.21***	1.89***	1.32***	1.78***	1.14***	.93***	1.30***	1.18**	1.42***	1.76***
6 Year of Degree	-.59***	-.30*	-.38**	-.37***	-.69***	-.24*	-.41***	-.02	-.59***	-.15	-.66***	-.19***
7 BA Prestigious	.24	.11	-.42	-.23	-.14	.15	-.18	.01	-.02	.13	.15	.45*
8 Graduate Prest.	-.24	-.86**	-.22	.36	.42	-.54	.68**	.21	-.27	.18	-.11	.03
9 Support	.31	.01	-.04	-.12	.00	-.11	.30-	.19	.15	-.22	.39-	.03
11 Years Academe	.86***	.46*	.28	.31**	.21	.36-	.74***	.35***	.77***	.71**	.11	.35***
12 Years Present	-.59***	-.30*	-.19	-.22-	-.11	-.14	-.34***	-.09	-.50***	-.32	-.34*	-.17*
13 Qual. Present	-.30-	.07	.04	.24	-.48	-.85**	.24	-.33*	.29	-.01	-.81**	-.18
14 No. Articles	1.28***	1.25***	.65*	.29-	.15	.42	.96***	.42*	1.19***	.31	.65**	.71***
15 No. Books	.35*	.51-	.32	-.06	.31	.21	1.01***	.64***	.12	.28	.64***	.09
16 Assoc.Research	.25	-.07	-.56	-.03	-.64	.26	-.29	-.46	-.30	-.18	-.36	-.10
17 No. Research	.01	.51*	.55**	.29	.11	.30	.08	.40*	-.02	.06	.02	.46***
18 No. Consulting	.84***	1.07**	.70***	.49***	.75**	.19	.27-	.44**	.45**	.21	.68***	.49***
19 Research/Teach	-.43**	.02	-.01	-.24	-.42-	-.03	-.27	-.16	-.49**	.23	-.21	.03
20 Administrative	.68***	.30*	.24*	.37***	.49***	.33-	.49***	.43***	.44***	.48**	.71***	.48***
21 Consulting	-.03	-.04	-.32	-.28*	-.24	.01	-.48**	-.06	.16	.15	-.37*	.02
22 Prof. Practice	-.27-	-.45	.31	-.09	-.37**	-.03	-.40*	-.47***	-.58**	-.03	-.78***	-.23*
23 Hours Taught	-.51***	-.02	-.35**	-.33***	-.21*	.08	-.36***	-.42***	-.51***	-.21	-.12	-.14***
24 Salary Base	2.48***	2.24***	2.87***	1.29***	1.14*	.71-	.96***	.02	2.11***	1.32**	2.71***	1.30***
27 Birth×No. Art.	-.15***	-.17***	-.07	-.03	.00	-.07	-.10***	-.01	-.11**	.03	-.05	-.07*
29 Birth×No.Child.	-.11*	-.18*	-.09	-.04	-.18*	.07	-.08-	.10-	-.14*	.07	-.08	.02
30 Sex×Mar.×Age	.31	.26	.22	-.07	-.07	-.14	.09	.00	-.15	.08	.08	.07
31 Part-time	-2.07***	-.96*	-.67	-1.61***	-2.57***	-.74	-1.12**	-1.87***	-1.50***	-.63	-1.72***	-.84***
No. Observations	1183	312	320	381	264	192	712	520	581	215	700	1029
No. Variables	26	26	26	26	26	26	26	26	26	26	26	26
Res'l d.f.	1156	285	293	354	237	165	685	493	554	188	673	1002
Multiple R-Squared	.67	.69	.67	.70	.68	.60	.75	.70	.71	.54	.58	.62
Res'l Mean Square	12.26	6.57	8.15	3.99	7.47	4.44	7.76	3.35	8.84	7.43	13.07	5.65
Mean Opp.Sex Res'l	3.47	-2.32	1.66	-1.07	2.28	-1.64	2.14	-1.29	2.91	-.94	2.29	-1.32
S.D. Opp.Sex Res'l	3.89	3.06	3.14	2.51	3.27	2.53	3.38	2.26	3.54	2.99	3.93	2.82

Table 2 (cont.). COEFFICIENTS OF THE MULTILINEAR REGRESSION EQUATION FOR PREDICTING FACULTY SALARIES (IN $1,000) IN RESEARCH UNIVERSITIES II AND DOCTORAL-GRANTING UNIVERSITIES I AND II

Field:	Bio/Phys Sci.		Education		Fine Arts		Humanities		Soc. Sci.		New Professions	
Predictors Sex:	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Constant	12.35	2.48	10.88	10.97	5.18	10.56	8.27	6.30	16.37	6.21	8.38	5.69
2 Date of Birth	.15	.25	- .22	- .30**	.30	- .43**	-.08	.03	- .05	.10	-.18	-.08
3 Marital Status	- .31	-.42	- .21	- .15	-.21	.31	1.00*	-.65*	- .12	-1.36^{-}	2.42**	-.84*
4 No. of Child.	.86***	1.34*	.43	- .88^{-}	.43	- .45	.45^{-}	-.28	.37	-1.25*	.43	-.09
5 Highest Degree	1.95***	1.62***	1.19***	1.23***	-.27	1.49***	.80**	.89***	1.97***	1.16**	2.03***	1.97***
6 Year of Degree	- .46***	-.15	- .29*	.06	.26*	- .05	-.08	.01	- .48***	.16	-.33**	-.04
7 BA Prestigious	.36	.40	.55	.14	.38	.98^{-}	.34	.27	- .70*	.11	-.58	.20
8 Graduate Prest.	.36^{-}	.19	.24	.08	.44	- .44	.59**	.27	.33	.32	.88**	-.14
9 Support	.14	-.07	.19	.27^{-}	.01	.23	-.04	.08	.36^{-}	-.34	.08	.01
11 Years Academe	.67***	.44*	- .06	.22*	.64***	.30	.98***	.27**	.53***	.80***	.16	.30***
12 Years Present	- .52***	-.24	- .07	- .06	-.08	- .06	-.38***	.11	-.61***	-.73***	-.10	-.03
13 Qual. Present	.01	.04	.17	- .12	-.05	- .64^{-}	-.05	.08	-.02	-.22	-.44^{-}	.06
14 No. Articles	.82***	1.04***	1.13***	.46*	.22	.14	.58***	1.38***	.52*	.94**	.64**	.24^{-}
15 No. Books	.49***	.85*	.59**	.08	.33	- .08	.63***	.04	.24	.08	.08	.41**
16 Assoc.Research	.13	.01	.16	- .55*	-.87	- .54	.49	.71^{-}	-.19	-.23	-.17	.30
17 No. Research	.01	.11	- .11	.29	.40	.92*	.03	.07	.11	.17	-.18	.39**
18 No. Consulting	.56***	.28	.43***	.18	.40^{-}	.38	.33^{-}	.41*	.21	.31	.86***	.64***
19 Research/Teach	- .32*	-.13	- .33	- .09	.13	.23	-.25	-.27*	.10	-.25	-.04	.02
20 Administrative	.65***	.58***	.32***	.17*	.52***	.36*	.30***	.39***	.44***	.45**	.72***	.39***
21 Consulting	.23	.09	- .21	- .11	-.24	.00	-.24	-.22**	-.09	.24	.12	-.04
22 Prof. Practice	- .58***	-.22	- .53**	- .34**	-.19	.03	.12	-.11	-.11	-.34	-.34*	-.38***
23 Hours Taught	- .32***	-.11	- .16^{-}	- .22***	-.10	.00	-.59***	-.16*	-.63***	-.14	.09	-.09*
24 Salary Base	1.86***	1.86***	2.97***	1.09***	1.68***	.71	.95***	.20	3.08***	.42	1.66***	1.33***
27 Birth×No. Art.	- .08**	-.09^{-}	- .13**	- .01	.02	.00	-.04	-.20***	-.05	-.09	-.04	.02
29 Birth×No.Child.	- .10*	-.24*	- .03	.10	-.07	.09	-.06	.04	-.02	.22*	-.09	-.03
30 Sex×Mar.×Age	- .14	.10	.19	.03	.19	- .28	-.21	.07	-.11	.42*	-.29	.16*
31 Part-time	-1.24***	-.97^{-}	- .46	-1.19**	-1.64*	.48	-1.89***	-1.44***	-1.77***	-1.12*	-.35	-.60**
No. Observations	941	254	368	468	241	204	602	553	499	189	500	883
No. Variables	26	26	26	26	26	26	26	26	26	26	26	26
Res'l d.f.	914	227	341	441	214	177	575	526	472	162	473	856
Multiple R-Squared	.66	.63	.65	.53	.56	.48	.72	.59	.67	.54	.58	.61
Res'l Mean Square	7.54	5.86	6.56	5.05	6.45	5.33	6.29	2.94	7.04	6.33	8.77	4.60
Mean Opp.Sex Res'l	1.56	-.58	2.26	-1.44	.23	-1.45	2.13	-.26	2.97	-1.59	2.54	-.13
S.D. Opp.Sex Res'l	2.86	2.49	3.01	2.72	2.81	2.64	3.07	2.13	3.42	3.18	3.21	2.54

Table 2 (cont.). COEFFICIENTS OF THE MULTILINEAR REGRESSION EQUATION FOR PREDICTING FACULTY SALARIES (IN $1,000) IN COMPREHENSIVE UNIVERSITIES AND COLLEGES I AND II

Field:	Bio/Phys Sci.		Education		Fine Arts		Humanities		Soc. Sci.	
Predictors Sex:	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Constant	11.08	5.59	18.55	6.75	16.58	8.74	18.71	10.63	21.04	7.57
2 Date of Birth	-.14	.02	-.30	.15	-.26	.30-	.01	.03	-.19	.21
3 Marital Status	.36	.42	.78	-.14	.72	-1.43-	.22	.25	1.12	-.44
4 No. of Child.	.06	-1.24	-.29	.45	-.49	-.62	1.81***	-.60	1.62*	-.74
5 Highest Degree	1.94***	1.26*	1.71***	2.21***	1.93**	.72-	.99**	1.55***	1.63**	1.50***
6 Year of Degree	-.22-	-.10	-.38**	-.08	-.25	-.13	-.40**	-.06	-.49*	-.16
7 BA Prestigious	-.43	.17	1.06	.63	1.10	.68	.25	.74-	-.08	-.35
8 Graduate Prest.	.61-	-.52	-.40	.50	.61	-.24	.06	.06	1.40*	.72
9 Support	.37	-.38	.24	.09	-.22	.16	.60**	.03	-.10	.15
11 Years Academe	.08	-.28	.36-	.28*	.45	.58*	-.00	.18	-.29	.24
12 Years Present	-.10	.63*	-.28	.11	-.05	.02	.38**	.44**	-.01	.36-
13 Qual. Present	-.75***	-.10	-.41**	-.41***	-.66**	-.17	-.35*	-.23*	-.70***	-.39*
14 No. Articles	1.57***	1.83***	.22	1.25***	-.08	.44	.15	1.01***	-.06	2.41***
15 No. Books	.75**	.30	.61-	.84**	.65-	2.04***	1.12***	.31	1.15**	.28
16 Assoc.Research	1.21*	1.54	-.89-	-.20	-.75	-.30	.06	-.89	-.18	.08
17 No. Research	-.47*	.54	.24	-.57	.38	1.79***	-.43	-.09	.25	.30
18 No. Consulting	.59*	-.66	.32	.25	.05	.54	.40	-.09	.37	-.12
19 Research/Teach	-.30	-.28	-.01	-.17	.02	.24	-.11	-.12	-.31	-.00
20 Administrative	.55***	1.13***	.29*	.22*	.70***	-.23	.00	.35**	.43*	.67***
21 Consulting	-.13	-.49	-.08	-.01	-.17	-.35*	-.50*	.14	.27	-.25
22 Prof. Practice	-.08	-.09	-.42-	-.28-	-.05	-.21	-.13	-.34-	-.29	-.25
23 Hours Taught	-.19-	-.28	-.23-	-.12	.13	-.15	-.68***	-.35**	-.57*	-.11
24 Salary Base	2.39***	2.03*	1.43**	1.51***	-.61	1.72***	.33	1.10***	.96	.54
27 BirthxNo. Art.	-.22***	-.19-	-.02	-.22**	.03	-.10	-.01	-.19***	.01	-.36***
29 BirthxNo.Child.	.03	.21	.07	-.13	.13	.05	-.26***	.10	-.26*	.14
30 SexxMar.xAge	-.31	.03	-.02	.07	-.39	.41*	-.07	-.01	-.01	.07
31 Part-time	-1.65**	-1.89*	-.76	-1.11*	-1.25	-2.56***	-1.80**	-1.75**	-1.20	-.99
No. Observations	253	124	194	303	123	153	238	347	177	139
No. Variables	26	26	26	26	26	26	26	26	26	26
Res'l d.f.	226	97	167	276	96	126	211	320	150	112
Multiple R-Squared	.75	.72	.65	.65	.61	.67	.69	.66	.64	.77
Res'l Mean Square	5.18	7.34	6.44	4.93	7.57	4.27	5.85	4.71	9.40	5.07
Mean Opp.Sex Res'l	.22	.43	1.36	-.28	1.98	1.31	.64	-.29	1.21	-.39
S.D. Opp.Sex Res'l	2.92	3.00	2.67	2.54	3.25	4.07	2.92	2.74	3.74	3.04

Table 2 (cont.). COEFFICIENTS OF THE MULTILINEAR REGRESSION EQUATION FOR PREDICTING FACULTY SALARIES (IN $1,000) IN LIBERAL ARTS COLLEGES I

Field:	Bio/Phys Sci.		Education		Fine Arts		Humanities		Soc. Sci.	
Predictors Sex:	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Constant	.35	6.17	10.13	4.78	11.89	11.01	19.81	15.43	8.04	13.03
2 Date of Birth	-.33	-.03	-1.22	.00	-.14	-.21	.06	-.13	.02	-.21
3 Marital Status	.62	-.40	4.02	-.33	-.34	-.55	1.25	-.19	1.55	-1.54-
4 No. of Child.	-.38	-.05	.23	-.53	.47	-.66	1.25**	.41	-.05	-1.45-
5 Highest Degree	.65	1.87***	-.44	.90*	.43	.12	.98**	.91***	.17	.36
6 Year of Degree	.52**	.11	.43	-.04	-.16	.13	-.61***	-.14	-.17	.20
7 BA Prestigious	-.07	-.80	3.12*	1.59**	1.74-	.81	.13	-.13	-.14	.26
8 Graduate Prest.	-.20	-.12	.58	.34	1.12	.40	.63-	-.21	.39	.45
9 Support	.42	.30	.76	.53*	.30	.06	-.11	.34-	.06	.52
11 Years Academe	.80***	.18	.20	.42*	.11	.23	.20	.60***	.69-	.07
12 Years Present	-.01	.56*	.04	-.26	.55	.05	.16	-.10	-.50	.00
13 Qual. Present	-.75**	-.48	-.54	.00	.40	-.46	-.81**	-.81***	-.60	-.70-
14 No. Articles	1.65***	.07	-1.00	.44	1.37-	.42	.28	.83***	.93-	.97-
15 No. Books	.18	1.27*	-.82	-.19	1.45*	.21	.83***	.02	.95*	.87
16 Assoc.Research	-1.16	-1.38	-2.56	1.46	-2.29	1.03	-1.47-	-.96	.40	-.42
17 No. Research	-.03	-.25	2.34	-.77	1.23*	-.15	-.07	.15	.54	-.01
18 No. Consulting	.70*	2.22***	-1.23-	.38	-.21	.43	.28	1.20*	-.02	.37
19 Research/Teach	-.19	.33	.04	.54	.49	-.17	-.17	.21	.71-	.07
20 Administrative	.15	.63**	-.20	.59***	-.32	.62*	.58***	.60***	.76**	.44-
21 Consulting	.24	-.17	.79	-.02	.21	-.21	-.17	-.34*	.01	-.14
22 Prof. Practice	.48-	-.99*	-.86	-.04	.69-	-.02	-.31	-.02	-.40	.30
23 Hours Taught	-.33**	.11	-.42	-.24*	-.63**	-.27	-.10	-.16	-.35	-.14
24 Salary Base	1.39**	-1.04*	3.06*	-.37	-1.48-	-.57	.55-	.58*	.65	-1.22*
27 BirthxNo. Art.	-.21***	-.02	.35	.00	-.28*	.04	-.02	-.12*	-.09	-.15
29 BirthxNo.Child.	.06	-.06	-.08	.06	-.16	.00	-.17*	-.06	.09	.19
30 SexxMar.xAge	.29	.17	.06	.17	.52	.15	-.16	-.26-	-.17	.26
31 Part-time	-.61	-.88	-.76	-1.76**	-.03	-2.02*	-.18	-1.65***	-1.59-	-1.82**
No. Observations	156	140	47	108	54	93	221	292	127	81
No. Variables	26	26	26	26	26	26	26	26	26	26
Res'l d.f.	129	113	20	81	27	66	194	265	100	54
Multiple R-Squared	.76	.75	.82	.75	.86	.67	.77	.69	.58	.78
Res'l Mean Square	3.82	4.86	5.18	2.44	3.55	4.35	4.65	4.13	7.81	2.25
Mean Opp. Sex Res'l	1.67	-1.68	2.17	-.67	1.97	-3.47	.23	.10	3.04	-.09
S.D. Opp. Sex Res'l	2.86	3.08	3.03	3.94	3.03	3.39	2.46	2.57	3.34	2.20

Table 2 (concl.). COEFFICIENTS OF THE MULTILINEAR REGRESSION EQUATION FOR PREDICTING FACULTY SALARIES (IN $1,000) IN LIBERAL ARTS COLLEGES II AND TWO-YEAR COLLEGES

Field:	Bio/Phys Sci.		Education		Fine Arts		Humanities		Soc. Sci.		New Professions	
Predictors Sex:	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Constant	-.90	5.37	9.56	2.80	-.79	3.83	-3.74	1.22	8.61	-4.49	17.41	-1.03
2 Date of Birth	-.11	.03	.30	.04	-.02	.43**	.33	.12	-.01	-.11	.36	-.09
3 Marital Status	.07	.20	1.44	-.06	-.80	-1.12	-.23	-.24	.15	.25	-2.85	-.04
4 No. of Child.	.54	3.03***	.56	.21	1.23	-.57	1.23**	.35	-.35	.03	.91	-.43
5 Highest Degree	.85*	.83	1.20	.08	1.88*	1.42**	.94**	1.04***	1.46*	.30	2.69**	.84*
6 Year of Degree	.15	-.02	-.20	-.03	.15	-.12	-.03	-.01	-.37	.43*	-.27	.03
7 BA prestigious	.58	1.37	.17	1.30*	-1.18	.13	.49	1.00**	-.57	-.36	-.50	1.09*
8 Graduate Prest.	.40	-.41	-.37	-.04	.19	-.44	.68	.52*	-.66	.26	-.248	.07
9 Support	.11	-.51	-.17	.74**	-.02	-.30	.05	.35*	.97*	.22	-.71	.27
11 Years Academe	.63***	.41	.69*	.29	.12	.30	.43*	.36**	.63	.45	.31	-.04
12 Years Present	-.09	-.26	-.55	-.04	.31	.32	.18	.11	-.17	.19	-.12	.27
13 Qual. Present	.41	.16	.38	1.08***	1.20*	.03	.59*	.49**	.95*	.61	-1.32	1.35***
14 No. Articles	-.24	.96*	-.28	.15	-.81	.76	.06	.17	1.19	.22	4.36	-.03
15 No. Books	.25	-.96	1.06*	-.33	.04	.18	.71*	.23	-.21	-.16	-1.28	.18
16 Assoc. Research	.17	-.31	-.65	-1.11	-1.32	-.56	1.36	.46	-1.59	1.27	2.77	-.53
17 No. Research	-.18	-.18	-1.46	-.56	1.59	.16	.45	.03	-.27	2.03*	.40	1.37*
18 No. Consulting	.19	2.01**	.45	.29	-.46	-.22	.52	.46	.40	.52	.36	-.36
19 Research/Teach	.33	-.27	-.31	.41	.34	.48	.28	.23	.12	.19	-.31	-.31
20 Administrative	.46***	.29	-.18	.06	.39	.26	.35*	.21*	.66*	.24	.19	.35**
21 Consulting	-.35	-.25	-.35	-.20	-.35	-.19	.13	-.11	-.57	.02	-1.59	-.07
22 Prof. Practice	-.52**	-.79*	-.02	-.36*	-.23	-.05	-.31	-.18	.03	-.65*	-.71	-.21
23 Hours Taught	-.07	.19	-.07	-.09	-.27	-.10	-.06	-.04	-.11	-.53*	.15	.02
24 Salary Base	.51	.13	1.08	.14	1.05	.13	-.17	-.05	1.34	.23	1.39	.52
27 Birth×No. Art.	.09	-.02	.04	.10	.27	-.19	.00	-.02	-.20	.05	-.53	.03
29 Birth×No.Child.	-.04	-.47***	-.11	-.17	-.13	.08	-.19*	-.08	.04	-.14	-.06	-.02
30 Sex×Mar.×Age	.26	-.20	-.17	.17	.21	.45*	.30	.09	.08	.27	.36	.29
31 Part-time	-.40	-.27	-1.88*	-.29	-.96	-.58	-1.37*	-1.38***	-.83	-1.60	-1.29	-.34
No. Observations	250	180	88	207	93	139	241	469	112	124	47	223
No. Variables	26	26	26	26	26	26	26	26	26	26	26	26
Res'l d.f.	223	153	61	180	66	112	214	442	85	97	20	196
Multiple R-Squared	.46	.36	.52	.32	.43	.48	.45	.39	.63	.48	.73	.42
Res'l Mean Sqyare	4.72	7.22	5.37	6.02	7.02	3.94	5.94	4.64	7.61	8.01	7.81	4.62
Mean Opp.Sex Res'l	-.06	-1.34	2.61	-.42	3.32	-1.99	1.39	-1.61	2.35	-1.65	2.61	-2.96
S.D. Opp.Sex Res'l	3.00	3.10	2.68	3.00	2.83	2.38	2.45	2.28	3.33	3.46	3.83	4.03

TABLE 3. CODED VALUES OF THE PREDICTOR VARIABLES FOR "TYPICAL" FACULTY MEMBER

Multipliers in the multilinear regression equation for predicting salaries.

Use corresponding number in parentheses for women. For predictor No. 6 (year of highest degree) in Biol/Physical Science, use 17 for faculty aged 40 years, use 15 for faculty aged 50 years.

	Type =	Research Univ. I		Res. Univ.II, Doc. - Grant. Univ. I, II		Comp. Coll., Univ. I, II		Lib. Arts Coll. I		Lib. Arts Coll. II, 2-Yr. Coll.	
	Age =	40	50	40	50	40	50	40	50	40	50
Predictors											
	Constant	1	1	1	1	1	1	1	1	1	1
2	Date of Birth	6	4	6	4	6	4	6	4	6	4
3	Marital Status	2	2	2	2	2	2	2	2	2	2
4	No. of Child.	3	3	3	3	3	3	3	3	3	3
5	Highest Degree	2	2	2	2	2	2	2	2	2	2
6	Yr. of Degree	18	16	18	16	18	16	18	16	18	16
7	BA Prestigious	1	1	1	1	1	1	1	1	1	1
8	Graduate Prest.	1	1	1	1	1	1	1	1	1	1
9	Support	1	1	1	1	1	1	1	1	1	1
11	Years Academe	4	6	4	6	4	6	4	6	4	6
12	Years Present	2	2	2	2	2	2	2	2	2	2
13	Qual. Present	1	1	3	3	2	2	1	1	3	3
14	No. Articles	4	5	2	2	1	1	4	4	1	1
15	No. Books	1	2	1	1	1	1	1	1	1	1
16	Assoc.Research	1	1	2	2	2	2	1	1	2	2
17	No. Research	1	1	0	0	0	0	1	1	0	0
18	No.Consulting	0	0	0	0	0	0	0	0	0	0
19	Research/Teach	2	2	3	3	4	4	2	2	4	4
20	Administrative	2	2	2	2	2	2	2	2	2	2
21	Consulting	2	2	2	2	2	2	2	2	2	2
22	Prof.Practice	1	1	1	1	1	1	1	1	1	1
23	Hours Taught	3	3	5	5	6	6	4	4	7	7
24	Salary Base	1	1	1	1	1	1	1	1	1	1
27	Birth×No. Art.	18	12	6	4	0	0	18	12	0	0
29	Birth×No.Child	6	4	6	4	6	4	6	4	6	4
30	Sex×Mar.×Age	4(8)	4(8)	4(8)	4(8)	4(8)	4(8)	4(8)	4(8)	4(8)	4(8)
31	Part-time/Full	1	1	1	1	1	1	1	1	1	1

are in less selective institutions.

We found that salaries are well predicted by the appropriate equation. Typical values of the multiple R^2 are above 0.6, or even 0.7, except for the last institutional type considered, a combination of Liberal Arts Colleges II and Two-year Colleges. Tests show these institutions to be somewhat heterogeneous, yet the sample sizes are too small for further subdivision. The value of R^2 tends to be smaller when the sample is small, which may be part of the reason the estimates for women often have smaller R^2 than the estimates for men in the same field and type. The standard deviation of an individual salary prediction is around $2,000 to $3,000 except for the Biological and Physical Sciences and the New Professions in the Research Universities I. Here, some men have considerably higher salaries than predicted.

3. COMPARISON OF SALARIES OF MEN AND WOMEN

Given the attributes of any individual, including sex and field and type of institution, we can use the appropriate set of coefficients shown in Table 2 with the predictor variables listed in Table 1 to estimate the corresponding faculty salary. Examples are worked out for a faculty member aged about 40 years in 1969 and for one aged about 50 years. We selected attributes which might be "typical" for these ages in the specified type of institution, making as little change

Table 4. ESTIMATED SALARY (IN $1,000) OF "TYPICAL" FACULTY MEMBER BY SEX AND AGE FOR VARIOUS FIELDS AND TYPES OF INSTITUTIONS

Note ·less than 100 residual degrees of freedom, :less than 60, ::less than 30

	Field = B/Phy.Sci.		Educ.		Fine Arts		Human.		Soc. Sci.		New Prof's	
	Age = 40	50	40	50	40	50	40	50	40	50	40	50
Type	Sex											
Research Univ. I	M	15.6	20.1	13.1	16.0	15.5	18.2	13.5	18.1	15.8	20.5	14.5	17.7
	W	12.5	16.3	12.3	14.4	9.9	12.2	11.5	13.6	12.4	14.7	12.7	15.0
Research Univ. II, Doctoral-Granting Univ.I,II	M	13.8	16.1	12.1	13.3	11.3	12.7	12.1	14.6	12.2	14.4	13.1	14.7
	W	12.6	14.0	9.7	10.4	9.3	10.7	9.7	10.4	11.2	12.1	11.2	12.1
Comp.Univ. and Coll. I and II	M	11.0	11.9	12.4	14.3	12.5·	14.2·	11.8	13.1	13.3	14.6	-	-
	W	7.8·	7.0·	12.8	13.5	10.3	11.1	10.3	10.5	9.8	9.9	-	-
Liberal Arts Coll. I	M	12.6	15.0	23.6::	23.7::	15.4::	18.2::	14.5	16.4	12.1	14.2	-	-
	W	9.6	10.1	10.3·	11.1·	11.2·	11.6·	10.9	13.5	9.6:	10.3:	-	-
Lib. Arts Coll. II, 2-Yr.Coll.	M	12.1	13.4	12.8·	14.2·	11.0·	11.2·	13.0	13.6	10.5·	12.4·	12.6::	13.2::
	W	13.2	15.0	10.9	11.8	9.3	9.1	11.7	12.3	9.2·	9.8·	9.9	10.0

as possible in the individual. First, these attributes were coded using Table 1 (see Bayer [7] for more details) with the results as listed in Table 3. Once the values of the predictor variables were fixed, each was multiplied by the appropriate coefficient in Table 2. The sum of these cross-products is the estimated salary for the individual, as shown in Table 4. When the sample used to estimate the coefficients was small, the resulting estimated salary is less reliable; these cases are marked by colons.

In almost every category, the predicted salary for men is larger than that for women. The differentials tend to be larger in the Research Universities I; indeed, whenever the predicted salary for men is large, the salary differential between men and women tends to be large. Also, the increase in salary from age 40 to age 50 is much less for women than for men, almost without exception. We see that women who have exactly the same attributes as men, that is, the attributes shown in Table 3, tend to be paid a much lower annual salary. The striking differential persists at other age levels and with any reasonable choice of attributes, tending to be more pronounced for older women. Recalling the very good fit of the multilinear regression equations to the actual salary, as evidenced by the high R^2, we must conclude that there is sex discrimination in faculty salaries and that it is especially strong in Research Universities I and for older women.

Our study of salary differences between men and women is retrospective, utilizing salaries and prediction equations of faculty actually employed in 1969. Whatever discrimination there may be that prevents the employment of women as faculty is not observed in the Carnegie Survey. We have no information on those who are highly qualified, as judged by our predictor variables, but were not employed as faculty in 1969. Further, as pointed out by Astin and Bayer [3], for those who were employed, we utilize the observed value of the predictor variables. If there has been any discrimination against women that affects the predictor variables, such as discrimination in graduate school that

makes it more difficult for women to obtain the doctorate, these effects are not taken into account. We presume that there is some such discrimination against women, at least in some of the predictor variables. Hence our estimated salary differentials, showing women receiving less than men, have probably been underestimated.

4. COMPARISONS OF ACTUAL SALARY WITH THAT PREDICTED FROM THE EQUATION FOR THE OPPOSITE SEX

In the last section we compared the estimated salaries of men and women faculty who have the same specified attributes and found large salary differentials due to sex, indicating discrimination against women. It is also of interest to examine the actual salary differentials in the various fields and types of institutions. Since, as noted at the outset, men and women faculty do not have the same distribution of the attributes we used as predictor variables, the difference between the actual salary a woman receives minus the estimated salary for a man with the same attributes in the same field and type of institution is an indication of the salary discrimination against women in that field and type of institution. We study the distribution of the residual: actual salary minus the estimate from the opposite-sex equation.

This residual was computed for each individual in the sample. The last two lines of Table 2 show the mean and standard deviation of the opposite-sex residual, separately for men and women, for each combination of field and type of institution. In Biological and Physical Sciences in the Research Universities I, men are overpaid, as estimated from the women's equation, by $3,470 on the average. Further, women tend to be underpaid, as estimated from the men's equation, by $2,320 annually. For almost every combination of field by type, when judged by the multilinear regression equation for predicting salary of the opposite sex, men tend to be overpaid and women tend to be underpaid. Moreover, noting the standard deviation, we see that the mean opposite-sex residuals are significantly different for men versus women.

The distribution of these residuals is approximately normal, with the mean for men shifted to the positive, the mean for women shifted to the negative. The distributions are plotted for six cases in Figure 1. In the more selective institutions, the shift between men and women is very large, also some men have extremely large positive residuals, when their salary is compared with the prediction from the women's equation. More than 80% of the women there are underpaid as judged from predictions for men of the same attributes in the same case of field by type. Instead of finding half of the residuals above zero and half below (as happens when comparing actual salary with own-sex equation), we find that the 50% mark is at about -$2,000 and that 20% of the women are underpaid by $4,000 or more annually.

When we shift attention to less prestigious institutions, the sex differentials are smaller. Nevertheless, some differential persists. Admittedly, the determination of salary is complex and the results presented are statistical. But the differences found are entirely too large to be due to chance and reflect discrimination. Indeed, as noted above, because of the probable sex bias in the predictor variables, we are probably underestimating the bias in the salaries of women.

One might argue that any particular institution is "different" from those of its type. With a little effort, each institution can compute the residual difference between actual salary and that predicted from the coefficients in Table 2 for each member of its faculty, using first the appropriate sex equation and then the opposite-sex equation. Each institution can thus compute two residuals for each faculty member, actual salary minus prediction from own-sex equation, and actual salary minus prediction from opposite-sex equation. We have done this for each institution in the Carnegie Survey in order to check that the grouping into types was satisfactory. Our computations show that the results are not very different from one institution to another.

The Carnegie data were collected in 1969 and one might hope that salary differentials due to sex are less pronounced now. The American Council on Education repeated the faculty survey in 1972-73. The data are not yet available in detail but Bayer [8] has already published extensive summary information from the survey. It is evident that there has been

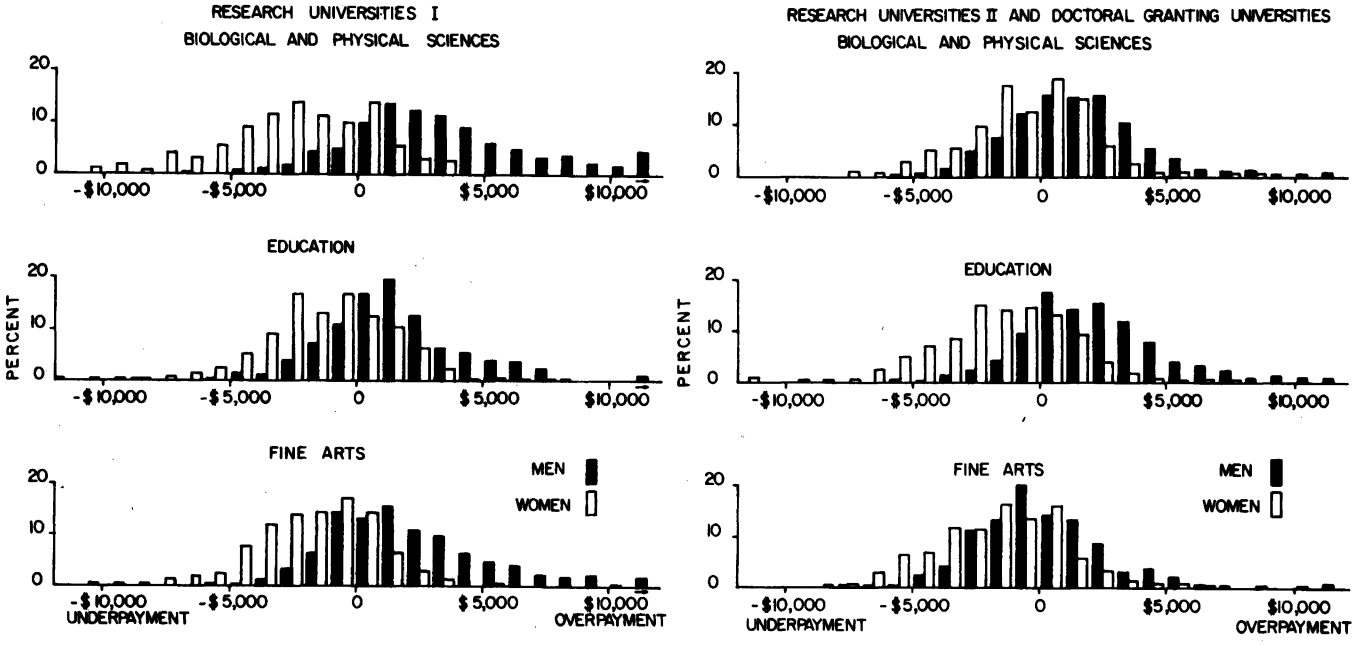


Fig. 1. Distribution of residual salary for men and women: difference between actual salary and that predicted from opposite-sex equation. Women tend to be underpaid compared with men of same ability and performance; men tend to be overpaid.

no appreciable change in the difference between the sexes.

5. DISCRIMINATION BY RACE AS WELL AS SEX

It is of interest to estimate the difference in faculty salary attributable to race as well as to sex. However, the numbers of nonwhite faculty in the Carnegie Survey are too small to allow a direct study by race, as was done for sex. But we can look at the residuals in salaries for each race separately. Do Black men tend to be underpaid when their actual salary is compared to the estimate computed from the equation for all men (and thus essentially for White men) in the same field and type of institution? Do Black women tend to be underpaid when their salary is estimated from the equation for all women in the same field and type? We can ask similar questions for Orientals and for Other races (Spanish surnames were not noted). Further, if we examine the distribution of residual salaries of Black men when their salary is estimated from the equation for all women, and similarly for other combinations of race and sex, we can determine whether race or sex is the more important in reducing salary.

For each combination of field and type of institution, we thus computed the salary residuals separately for each combination of race (White, Black, Oriental, and Other) with sex (men, women). The results are clear and consistent even with the irregularities to be expected from very small samples. There is a slight tendency for Black men to be underpaid when compared with the all men predictions; there is a slight tendency for Black women to be overpaid when compared with the all women predictions. Similar differentials appear for Orientals but more irregularly. But these differences are not significant. On the other hand, the differences due to sex persist and are strong; not only do White men tend to be overpaid as judged by the all-women prediction equations, so do Black men, so do Oriental men, and so do Other men, all by about the same mean amount. Conversely, as judged by the prediction equations for all men, women tend to be underpaid in almost every combination of field and type, no matter what race they may be.

The apparent discrimination in faculty salary due to sex is strong and persists for every race. Perhaps surprisingly, the apparent discrimination due to race is small; it is not significant. We are probably underestimating the discrimination due to race even more than we are underestimating that due to sex, considering the additional discrimination in the predictor variables such as access to graduate school or access to academe itself.

6. COMPARISON WITH OTHER STUDIES

There have been several studies of the salary differences between men and women faculty that consider the possibly different distribution of attributes and performance between the two groups. Most studies have concentrated on a single institution or a small group of universities and have restricted attention to a few departments. Insofar as their results overlap this paper; the agreement is good.

The initial results of an extensive study by Johnson and Stafford are reported [9] by Committee Z of the American Association of University Professors. The data used are taken from the 1964 and 1970 National Register collected by the National Science Foundation, so that disciplines in the humanities and professions are not included. Further, productivity is not measured. The preliminary results for five departments show percentage losses in salary of women consistent with our results; they note also that the salary differential increases sharply with increasing time from the Ph.D. They question whether the differential may in part be due to women's tendency to withdraw from the labor market during child-bearing years, something on which they do not have data. The facts are that highly educated women do not withdraw from the labor force more than men. The latest ACE survey found [8] that nearly one-fourth of all faculty had interrupted their professional careers more than one year and that, moreover, a greater percentage of men than women had done so. This is another indication that one should not try to blame salary differentials on unobserved variables.

REFERENCES

[1] National Education Association, Research, Salaries Paid and Salary Related Practices in Higher Education, 1971-72, Washington, D. C., 1972.

[2] Trow, Martin, Carnegie Commission National Survey of Higher Education, Technical Report, Carnegie Commission on Higher Education, Berkeley, California, 1972.

[3] Astin, H. S. and Bayer, A. E., "Sex discrimination in academe," Educational Record, 53 (1972), 101-118.

[4] Carnegie Commission on Higher Education, Opportunities for Women in Higher Education: Their Current Participation, Prospects for the Future, and Recommendations for Action, New York: McGraw-Hill, 1973.

[5] Available through VIM Library, Software Distribution Department, Control Data Corporation, 3145 Porter Drive, Palo Alto, California 94304.

[6] Dixon, W. J., editor, BMD Biomedical Computer Programs, Berkeley and Los Angeles: University of California Press, 3rd ed., 1973.

[7] Bayer, A. E., College and University Faculty: A Statistical Description, ACE Research Reports, 5, No. 5, Washington: American Council on Education, 1970.

[8] Bayer, A. E., Teaching Faculty in Academe: 1972-73, ACE Research Reports, 8, No. 2, Washington: American Council on Education, 1973.

[9] Steiner, P. O., Eymonerie, M., and Woolf, W. B., "Surviving the seventies: Report on the economic status of the profession," AAUP Bulletin, 59 (1973), 188-258 (especially 204-205).

This paper was prepared with the partial support of NIH Research Grant No. GM-10525.

MEASUREMENT OF EARNINGS DIFFERENTIALS BETWEEN THE SEXES

Joseph L. Gastwirth, The George Washington University

Introduction

Anyone looking at earnings data is immediately struck by the difference in the average earnings of men and women. Since the principles of our country state that pay should be based on merit and skill factors and that everyone doing the same job should be paid the same wages, explaining the observed wage differences is a high priority social problem.

Our research began when I served as a Visiting Faculty Advisor to OMB and participated in a task force under the direction of the Deputy Assistant Attorney General for Civil Rights. While a substantial literature has been devoted to discovering the factors contributing to high earnings power (e.g., education, experience) and measuring (via regression models) their relative importance, it appeared that one needed a simple, yet statistically sound measure that would 1) enable one to detect areas of the labor market in which women are furthest behind men 2) be applicable to regularly issued statistical series so that progress could be followed over time.

2. Measures of Differentials

The problem of comparing male and female earnings distributions can be regarded as a two-sample problem. We consider the wages of women and men to come from theoretical distributions F(x) and G(x), respectively (where F(x) denotes the fraction of women earning less than x). The Census Bureau often uses the ratio of the medians to compare earnings and income distributions. Recently, they [9] have considered a new Overlap measure. In this section we review the Overlap measure and introduce a Probability measure based on the Wilcoxon test which we feel is superior for the current purpose.

The overlap measure (OVL) is best described in terms of the density functions f(x) and g(x) corresponding to F(x) and G(x). It is defined as the area under both g(x) and g(x), i.e.,

$$OVL = \int_0^\infty \min(f,g)\,dx, \qquad (2.1)$$

and is the shaded area in Figure 1.

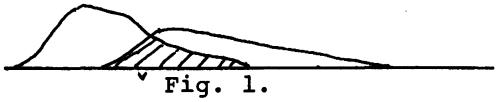


Fig. 1.

While the overlap measure has a nice pictorial representation it has several drawbacks: 1) Two widely different pairs of distributions can have the same value. Moreover, the cause of the disparity in each pair of distributions can be different. This can be illustrated by 2 pairs of distributions with OVL = 0.

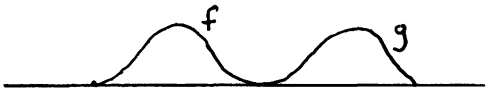


Fig. 2.a.

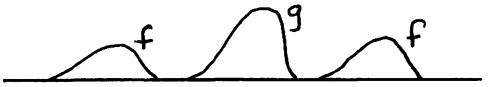


Fig. 2.b.

In Figure 2a, every member of the G population earns more than any member of the F group. In Figure 2b both populations have the same mean but the G population is concentrated near this mean value while the F distribution is really composed of two separate groups. 2) The OVL measure places undo-emphasis on the point, v. (see Fig. 1), where the density functions intersect. In particular, if a female earning more than v obtains a pay raise (while everyone else in both populations remains at the same level) the value of OVL remains unchanged.

As the second objection to the overlap measure also applies to the ratio of medians, we propose to use the probability that a randomly selected woman earns at least as much as a (randomly chosen) man. In order to rank industries, however, general social phenomena which depress women's wages relative to men tend to operate "across the board" so that the PROB measure should serve our purpose. More importantly, the PROB measure will detect any advancement of women relative to men so that it can be used to "monitor" upgrading programs.

Mathematically, the probability that a woman earns at least as much as a man can be expressed in terms of the distributions F(x) and G(x) as

$$PROB = \int_0^\infty [1-F(x)]g(x)\,dx. \qquad (2.2)$$

Because the PROB measure is related to the Mann-Whitney form of the Wilcoxon test, its standard deviation is known [4].

3. Analysis of the Longitudinal Social Security Data

The finding that women do not receive the same economic rewards as men for continuous labor force participation hardly needs extensive statistical documentation. Our task, however, is to show that the PROB measure reveals this fact and also detects small changes. Thus, one can use it to track the status of women over time.

The data base consisted of earnings data for two time periods, e.g. 1965-1970. Workers in the last period (1970) are split into those who worked in 1965 and

those who did not. (This group of workers consists of new and re-entrants to the labor force.) Similarly, the workers in the earlier period are grouped into those who had earnings in 1970 and those who did not. In order to avoid confounding differences based on sex with those due to race we discuss the results for the white population.

In Table 3.1 we present the values of the various measures for the total U.S. for the 1965-70 and 1962-67 periods. The earnings used were reported in the first quarter (multiplied by four).

While all three measures show a great disparity in the male-female earnings differentials between workers who entered (or re-entered) the work force in the 5 year period before 1970 (or 1967) and those who had worked five years earlier, it is more enlightening to look at men and women who worked at both times. When we contrast the differentials based on 1970 earnings to those derived from the 1965 earnings both the PROB and OVL measures DECREASED while the ratio of median incomes INCREASED. The same phenomenon also held during the 1962-67.

4. The Relative Status of Women in Various Industries

In order to demonstrate the utility of the PROB measure we apply it to Social Security data (by industry) for 1966.

Before presenting the results, some technical limitations of the data should be noted. The data is based on a 1% sample of earning records of people who worked in all four quarters. The data cannot distinguish between full and part-time employment, however, so that the over-all status of women may be biased downwards. The advantage of the Social Security data is that the sample is extremely large (390,000).

In Table 4.1 we present our results. Several major features emerge: 1) In almost all industries, black women fare better relative to black men than white women do to white men. 2) All the measures show that, across all industries, women to not fare well, however, the relative rankings do not agree. In particular, the Probability and Overlap measures give low scores to Communications, Public Utilities and Manufacturing (which, unfortunately, is highly aggregated here) while the Ratio of Medians yields a low rank for Retail Trade, Manufacturing, Communications and Services and a high one for Public Utilities and Transportation. 3) The PROB measure generally is relatively further away from its "ideal" value, 1/2, than the other measures are to theirs. This is a desirable property for tracking purposes.

5. Analysis of Occupational Data within an Industry

The low values of all the measures of earnings differentials in the aggregate industry data presented in section 4 can result from a) paying women less than men to for the same job and/or b) excluding women from jobs on a career ladder leading to promotions etc. thereby clustering women in the relatively low paying positions. The longitudinal study in section 3 supports the second explanation. The only available data enabling us to shed some light on this question is the BLS area Wage Survey which is collected regularly from firms on wages for precisely defined jobs.

In Table 5.1 the PROB value is given for a variety of Professional and Office jobs. The two sets of numbers tell different stories. Generally, earnings are nearer "equality" in the Professional categories than in Office jobs. Since women dominate (numerically) the Office jobs surveyed these results cannot be explained by excuses such as lack of qualified applicants with relevant work experience etc. Moreover, the higher the proportion of women in an occupation, the lower is their probability of equal pay. For example, females outnumber male Billers by 8:1 and PROB = .17,[1] they outnumber male order clerks 5:2 and PROB = .19. In the occupations with skill categories, e.g., Accounting Clerks - at the highest level women outnumber men 5:2 and PROB = .25 while at the lower level (B) women outnumber men by about 6:1 and in all industries (except Manufacturing) the PROB value is lower.

The data for Tabulating Machine Operators illustrates the relationship between employment segregation and lower pay for women. For the highest skill level (A), where men outnumber women 5:2, PROB = .35 and the ratio of medians equals .90. For the lower skill levels (B and C) where men still outnumber women the results are similar. Only in the Public Utilities (level B) workers where women outnumber men 2:1 does PROB fall to .26 and the ratio to .80.

Looking back at Table 5.1, in this light, one wonders whether the relatively high values of all our measures give most Professional and Technical occupations results from the scarcity of women in them. The only occupation in an industry category where women are employed in nearly the same numbers as men was computer operators in the Public Utilities[2] - which received the lowest PROB score in the Professional class.

6. Summary

The purpose of our paper was to illustrate how a simple measure of earnings differentials can be used to rank industries (or occupations) and to monitor progress over time. By analyzing several

U.S. government data series we showed that

1) Women do not receive the same economic return for continuous work as men. Indeed, they fall further behind as time (in the labor force) passes.

2) The relative status of the sexes is nearer equality in occupations in which men are employed in substantial numbers. Low values of our measure of equality occurred where women dominate (numerically) the job.

2) In order to obtain a complete picture one should study employment as well as earnings data. A low score can result from a variety of factors, e.g. new hiring as well as placing women primarily in low paying jobs.

It is important to remember that broad statistical series cannot prove that discrimination exists, however, the tools developed can aid in the process of monitoring progress. Moreover, the data used in the section 5 is available to most large companies. If the Public Utilities data had been for one firm, management could immediately spot that something might be amiss in their computer-operator division.

In addition, I hope this paper will also stimulate professional statisticians to work with various government data series and point out to other social scientists who use the data which data sets are most appropriate for various types of analysis. Finally, I would like to thank the Women's Caucus of ASA for honoring me with the invitation to prepare this paper.

Acknowledgement: It is a pleasure to thank David Melcovsky for his assistance with the computer aspects of the paper and Carol Fey for typing several versions. Finally, the partial support of this research by an NSF institutional grant to George Washington University is gratefully acknowledged.

FOOTNOTES

[1]Notice, however, in the Public Utilities where the number of female and male Billers are nearly equal the PROB = .35.

[2]In the Area Wage Survey Public Utilities includes Utility Companies and the Communications and Transportation industries.

REFERENCES

[1] Buckley, J. E. "Pay Differences between men and women in the same job" Monthly Labor Review Nov. 1971 pp. 26-39.

[2] Fuchs, V. R. "Differences in hourly earnings between men and women" Monthly Labor Review May 1971 pp. 9-15.

[3] Gastwirth, J. L. A New Index of Income Inequaltiy. Proc. I.S.I. meeting in Vienna, 1973.

[4] Gibbons, J. D. Nonparametric Statistical Inference, McGraw Hill: New York, 1970.

[5] Henle, P. "Exploring the Distribution of Earned Income," Monthly Labor Review Dec. 1972 p. 16-27.

[6] Kreps, J. Sex in the Marketplace: American Women at Work Johns Hopkins Press. Baltimore, Maryland, 1971.

[7] McNulty, D. J. "Differences in Pay Between Men and Women Workers" Monthly Labor Review Dec. 1967 p. 40-43.

[8] Strasser, A. Differentials and Overlaps in Annual Earning of Blacks and Whites Monthly Labor Review Dec. 1971 p. 16-26.

[9] Weitzman, M. Measures of Overlap of Income Distributions of White and Negro Families in the U.S. Tech. Paper 22 Bureau of the Census, 1970.

Table 3.1 Measures of Male-Female Differences for the United States White Population

	PROB	OVL	RATIO	MEDIAN MALE	MEDIAN FEMALE	MEAN MALE	MEAN FEMALE
1970							
All Workers	0.255	0.569	0.505	7454	3764	8027	4024
Worked in 1965	0.217	0.530	0.540	8424	4548	9224	4850
New Workers	0.427	0.819	0.839	3369	2825	4370	3062
1965							
All Workers	0.243	0.560	0.507	5623	2853	5988	3020
Worked in 1970	0.234	0.547	0.527	5879	3097	6266	3261
Dropouts	0.309	0.658	0.536	4380	2349	4880	2561
1967							
All Workers	0.275	0.606	0.417	5590	2331	6070	2721
Worked in 1962	0.222	0.536	0.498	6761	3370	7380	3532
New Workers	0.446	0.867	0.795	1492	1186	2408	1733
1962							
All Workers	0.266	0.604	0.427	4599	1965	5074	2255
Worked in 1967	0.264	0.604	0.478	4866	2325	5313	2520
Dropouts	0.330	0.711	0.424	2831	1200	3834	1669

Source: Bureau of Economic Analysis (Regional Econ. Div.) Department of Commerce

Table 4.1 Measures of the Relative Status of Women in Various Industries. Derived from the 1966 Social Security Data for 4-Quarter Workers

	White			Black		
Ind.	PROB	OVL	RATIO	PROB	OVL	RATIO
Total	.185	.480	.508	.252	.620	.570
Construction	.217	.545	.570	.312	.648	.717
Mining	.211	.567	.663*	.162	.204	.407*
Manuf.	.136	.404	.526	.226	.567	.606
Trans.	.205	.513	.710	.335	.674	.870*
Commun.	.105	.293	.523	.342	.520	.816
Pub. Util.	.150	.434	.637	.337	.431	.861
Whol. Tr.	.178	.472	.574	.275	.631	.691
Ret. Tr.	.229	.521	.480	.322	.693	.709
Finance etc.	.176	.426	.564	.410	.830	.914*
Service	.259	.612	.534	.308	.697	.599

*These values are based on a very small sample.

Table 5.1 The PROB Measure of Earnings Differentials Evaluated on Weekly Wage Data (1970-1971)

Prof. & Tech. Occupations	Industry	PROB	MEDIAN FEMALE	MEDIAN FEMALE	TOTAL FEMALE	TOTAL MALE
Comp Op. A	All	0.400	158.39	166.72	950	10834
	Manufacturing	0.464	165.57	168.58	383	4830
	Finance	0.356	146.05	158.39	287	2738
Comp Op. B	All	0.299	125.35	144.10	3984	18801
	Manufacturing	0.360	136.16	148.38	1176	6996
	Public Utilities	0.093	115.22	165.21	982	1212
	Wholesale Trade	0.300	126.82	147.29	424	2002
	Finance	0.313	120.40	134.43	976	5494
Comp Op. C	All	0.371	110.87	120.58	2634	7957
	Manufacturing	0.394	117.32	126.50	766	2439
	Public Utilities	0.243	105.21	130.62	795	498
	Finance	0.378	104.20	113.92	577	3058
Comp Prog A	All	0.424	217.88	226.72	1991	11131
	Manufacturing	0.411	219.62	230.42	558	4836
	Public Utilities	0.340	210.68	233.42	269	974
	Finance	0.464	210.39	215.56	607	2682
	Services	0.437	219.86	224.84	283	1303
Comp Prog B	All	0.430	182.39	189.89	3901	13987
	Manufacturing	0.439	187.48	195.02	1136	5575
	Public Utilities	0.412	195.59	204.29	481	1503
	Finance	0.453	175.84	179.22	1655	4220
Comp Prog C	All	0.443	156.61	161.58	2323	5747
	Manufacturing	0.413	159.76	170.52	632	2003
	Public Utilities	0.470	169.73	172.05	308	637
	Finance	0.474	150.97	152.83	1037	2190
Office Occupations						
Billers, Machine	All	0.173	100.64	153.80	8695	1159
	Public Utilities	0.350	143.40	162.01	1207	940
Clerks, Acct., A	All	0.255	127.66	152.19	54544	19228
	Manufacturing	0.258	132.51	156.20	20641	9326
	Public Utilities	0.246	130.56	162.70	7314	3159
	Wholesale Trade	0.268	128.95	150.04	6343	2389
	Retail Trade	0.295	118.94	138.52	6561	804
	Finances	0.265	117.29	138.24	9044	2295
	Services	0.355	130.35	142.73	4654	799
Clerks, Acct., B	All	0.218	100.42	127.92	103110	10211
	Manufacturing	0.253	105.62	127.86	30861	3399
	Public Utilities	0.191	105.83	142.73	16292	2413
	Wholesale Trade	0.203	103.14	133.00	13813	1950
	Retail Trade	0.363	94.48	105.75	18130	443
	Finance	0.201	93.01	114.03	17285	1467
	Services	0.312	101.96	114.21	6733	546
Secretaries, C	All	0.248	138.02	162.50	98895	368
Tab.Mach. Op. A	All	0.345	142.17	157.30	1002	2583
Tab. Mach. Op. B	All	0.355	117.46	130.60	3199	3911
	Public Utilities	0.261	113.81	143.61	986	484
Tab. Mach. Op. C	All	0.352	99.85	111.54	1805	1861

STUDYING CHANGE IN SEX-ROLE DEFINITIONS VIA ATTITUDE DATA

Karen Oppenheim Mason, University of Michigan

Sex-role definitions refer to social expectations for and beliefs about the behavior of women and men which members of a society agree upon. These definitions have both obvious and subtle influences on equality of opportunity for women. Public support for working women's rights to equal work or pay can form a "climate of opinion" within which supervisors, employers, the courts and legislators operate. Beliefs about the "natural" traits and competencies of women can influence how they are evaluated in the labor market or polity. Traditional norms which make women's career interests subordinate to their husbands' can influence women's general labor market opportunities; for example, by restricting or increasing their ability to move in order to find better employment.

Although social scientists have devised a number of ways in which to measure such social norms and beliefs [e.g., 6], sex-role attitude items employed on sample surveys provide one simple way of doing this. The study of such attitudes, viewed as measures of sex-role definitions, is of especial interest at this point in history. The recent rise of several sex-role liberation movements[1] has been much publicized in the media, provoking considerable controversy over the status and roles of women and men. However, to what extent these movements have brought about or been accompanied by change in sex-role expectations and beliefs of the general population (not just certain elites) remains unclear. Since these movements might be expected to influence publicly-expressed attitudes more quickly than they influence the structures which determine women's status attainment opportunities, a reading of these movements' impact on sex-role attitudes is obviously of interest at a time when such movements have become so highly active.

Table 1 displays some typical sex-role attitude items employed in national polls and other national surveys in the past. Such items admittedly provide what is from many points of view a superficial measurement of individuals' attitudes. In most instances, they make no attempt to measure the salience of particular opinions or beliefs, the intensity with which these are held, the extent to which they are integrated into self-conscious ideological systems, or the extent to which they dispose individuals to behave in particular ways. Such items, however, can nonetheless provide important clues about change in sex-role definitions. As the study of similar attitudes towards blacks among the white population over the past three decades suggests [7, 11], observed changes in responses to such survey items can at least indicate whether it has become unfashionable or frowned upon to display traditional prejudices and stereotypes; a change which can in turn pressage change in individuals' more deeply felt attitudes about the sexes and their social roles.

TABLE 1. Simple Percentages Agreeing with Fifteen Sex-Role Attitude Items in Four U.S. Sample Surveys. [Note: Some figures are preliminary.]

Attitude Item	1964 NORC College Seniors Study[a] Men	1964 NORC College Seniors Study[a] Women	1970 15-Year Follow-up Study[b] (Women Only)	1970 National Fertility Survey[c] (Women Only)	1973 North Carolina Study[d] (Women Only)
It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family.	e	e	--	78	65
It is more important for a wife to help her husband[s career]* than to have [one]* a career herself.	76	76	81	--	68
If a wife earns more money than her husband, the marriage is headed for trouble.	43	44	--	--	39
Men should share the work around the house with women such as doing dishes, cleaning and so forth.	--	--	--	53	75
Women who do not want at least one child are being selfish.	--	--	30	--	19
A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.	35	54	--	47	55

Item	1964 Study Men	1964 Study Women	1970 Follow-up (Women Only)	1970 National (Women Only)	1973 Study (Women Only)
A pre-school child is likely to suffer [emotional damage]* if his mother works.	66	60	--	71	49
A woman can live a full and happy life without marrying.	--	--	60	--	64
A man can make long range plans for his life, but a woman has to take things as they come.	39	38	--	31	--
Parents should encourage just as much independence in their daughters as in their sons.	66	82	87	--	--
Men and women should be paid the same money if they do the same work.	--	--	--	95	97
A woman should have exactly the same job opportunities as a man.	--	--	--	63	78
Women should be considered as seriously as men for jobs as executives or politicians or even President.	--	--	--	55	71
A woman's job should be kept for her when she is having a baby.	--	--	--	58	81
Women should stop expecting special privileges because of their sex.	e	e	--	--	43

*Words used only in 1964 version of the question.

[a]Based on a national probability sample of graduating arts and science college seniors of June, 1961, who were followed-up by mail questionnaire in 1964. The sample contains very few nonwhites. Most persons in the sample were age 24 in 1964.

[b]The initial sample for this study was a probability sample of high school sophomores and seniors residing in small and medium-size metropolitan areas as of 1955. These individuals were followed-up by mail questionnaire in 1970. This sample also includes very few nonwhites.

[c]Based on a national probability sample of ever-married women under age 45 in 1970. Information was collected through household interviews. This sample contains a representative proportion of blacks and other nonwhites, and is far more hetereogeneous with respect to educational attainment than are the other three samples.

[d]Based on a quota sample of women residing in three North Carolina metropolitan areas who were first married in 1962-1964 and were married only once and still married (husband present) as of 1973. Information was collected through household interviews; approximately one quarter of the sample is black. Over 90 percent of the women in this sample fall into the 25-34 age range; their educational attainment is somewhat higher than is that of the comparable age group in the 1970 Census.

[e]Item asked in 1964, but marginals currently unavailable.

As Duncan [3] has noted, the basic strategy involved in using survey measures of this type for assessing social change is replication of the items over time. In order for this replication to yield valid estimates of change, however, at least three conditions must be met. First, the replication of the items, their wording, response categories and coding, must be precise. As several studies [4, 10] have shown, even what appear to be minor variations here can seriously alter the stimulus presented to respondents and thereby change response distributions when underlying opinions are constant.

Secondly, the cross-sections on which these items are measured must be relatively general and "representative" of a society's population, and must also be well-defined enough to permit the matching of these cross-sections over time. This latter requirement is especially important when one is interested in estimating the potential effects of social movements on attitudes rather than assessing change attributable to shifts in social and economic characteristics of the population,

such as educational attainment, church attendance or labor-force participation rates.

Finally, if survey items are to be used to study sex-role definitions (not just particular sex-role attitudes or opinions), each cross-section studied must include several items, tapping a variety of aspects of these definitions. Time series for single items, such as Gallup's measuring approval of married women working [5], while of interst, tell us less about overall shifts in sex-role definitions than do series in which attitudes towards many aspects of these definitions are measured.[2]

The analyst concerned with measuring change in sex-role definitions over time can, in future years, create a time series meeting these three requirements by drawing periodic national probability samples and administering the same attitude items to each. Aside from the high costs of such an enterprise, however, there are two reasons for first seeking already collected and available data which at least in part meet these criteria. First, by analyzing attitude items from past surveys in terms of individuals' social backgrounds and behavior, we may better understand what these items generally are measuring; information that would be useful before we spend considerable resources on creating future series of national probability samples. In addition, previously collected surveys are also of interest because of the possibility that history may already have passed us by. While it seems dubious that the sex-role liberation movements have already wrought extensive and dramatic shifts in sex-role definitions, this possibility suggests it is important to assess change occurring in the late 1960's and early 1970's, as well as in the future.

The already collected and available data containing measures of sex-role attitudes of which this author is aware leave something to be desired in terms of the three requirements mentioned earlier for using such data to study change associated with social movements. However, four surveys, shown in Table 1, conform with these requirements enough to make analysis of them at least of interest, even if not definitive. Three of these four surveys employ probability samples which are national in scope (although in all three the sample universe is much more restricted than is ideally desirable). In addition, among the four there are 15 items which are replicated, in most instances precisely, these 15 covering a variety of aspects of sex-role definitions. With proper adjustment for differences among these surveys in sample universes and sample composition, then, we might obtain at least prima facie evidence concerning recent change in sex-role definitions. I will describe briefly in the remainder of this paper how this adjustment might be done.

We know from past empirical investigations of sex-role attitudes in the United States [8, 9, 13] that several individual traits, experiences and characteristics tend to correlate with sex-role attitudes. Among younger American women studied in 1970, for example, educational attainment, work experience, religiosity and race all have rather marked relationships with responses to several of the items shown in Table 1 [9]. The simple percentages agreeing with the 15 items shown in this table thus may differ across surveys for two reasons; either because there have been genuine historical changes in how people feel about the sexes and their roles, or because these samples vary on those social and economic traits which correlate with individual's sex-role attitudes. The latter possibility must be entertained in light of known differences among the four studies in their social and economic composition (see table footnotes).

Two statistical procedures are suggested for obtaining estimates across surveys which are interperable in terms of recent attitude change accompanying the rise of social movements. First, for individual characteristics which are constants in some surveys because of restrictions in sample universes, but which are variables in other surveys, a matching procedure in which analysis is limited to observations in each survey having the same range of characteristics is suggested. For example, in comparing responses from the 1964 study of college graduates with those from later surveys with more general samples, we would want to limit analysis of the latter to only those people who have at a minimum also graduated from college. Such matching does, of course, restrict the generality of the universe for which we are investigating change in sex-role attitudes, but this restriction is important if we are going to be able to interpret differences among surveys in terms of social change.

Simple case selection or matching procedures of this type are likely to remove gross compositional differences between surveys, but even after such procedures are used, further compositional differences for characteristics on which there is internal variation within each cross-section are likely to remain. Adjustment of attitude responses for these further compositional differences can be achieved in several ways, using available methods. The percentages agreeing with particular items in each matched subsample can, for example, be standardized for other social and economic variables, either directly or indirectly [1]. While an estimate of the magnitude and direction of change in attitudes over time can be made by comparing such standardized percentages, this approach normally does not yield estimates of the effects of the compositional variables on the attitudes; estimates which may be of interest if the compositional variables have themselves changed in the general population over time (as is, for example, the case for women's labor-force participation rates).

An alternative approach which both standardizes the time comparison for other social and economic traits of individuals, and which also yields direct estimates of the effects of these traits on individual's attitudes, consists of estimating a simple regression model of the following form for a sample consisting of the two matched cross-sections we are comparing:

$$A = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n + d\,T + e.$$

Here, A is a dummy variable measuring whether the individual agrees with a given attitude item or

not; a is the equation intercept; the b_i are partial coefficients associated with various individual social and economic traits measured by the X_i; and d is the partial coefficient for a dummy variable, T, which measures whether an observation falls into the later cross-section or the earlier one (e is the error term).[3] The direction and magnitude of change in attitudes over time not directly attributable to individual traits is estimated by the coefficient, d. If the sex-role liberation movements are being accompanied by change in public sex-role definitions consistent with their ideological thrust, then we would expect this coefficient to have a sign indicative of more egalitarian sex-role outlooks in the later cross-sections than in the earlier ones.

How well this procedure will suffice for estimating changes over time depends not just on the quality of the original data, but also on how well the regression model used to estimate change is specified. If our understanding of the determinants of individual attitudes suggests to us that our regressions include all relevant X_i, then we will be able to interpret the coefficient, d, in terms of historical change with some confidence. If, however, we are uncertain as to what particular X_i belong in our model, or have adequate measures of only some of the X_i we think are relevant, then our ability to interpret this coefficient in terms of recent social change will be considerably lessened. Thus, as is also the case in studying change in other forms of social behavior and ideology [2], the building of a theory or causal model of sex-role attitudes and definitions is an important accompanyment to the analysis of social change. The statistical procedures involved in matching or in regression analysis will take us only as far as our substantive understanding of people's attitudes will allow.

FOOTNOTES

[1] We include here the gay and men's liberation movements, as well as the women's liberation movement.

[2] The availability of many items tapping different aspects of sex-role definitions is in part important because such items can be analyzed within each cross-section for underlying dimensions, internal consistency and other characteristics indicating the structure and nature of these definitions.

[3] Because the dependent variable in this example can take on only two values, techniques such as logit analysis [12] would be more appropriate for estimating the parameters of the model than would ordinary least squares.

REFERENCES

[1] Barclay, George W., Techniques of Population Analysis. New York: John Wiley & Sons. 1958.

[2] Duncan, Otis Dudley, "Discrimination Against Negroes," The Annals of the American Academy of Political and Social Science 371(May):85-103. 1967.

[3] _____, Toward Social Reporting: Next Steps. New York: Russell Sage Foundation. 1969.

[4] _____, Howard Schuman and Beverly Duncan, Social Change in a Metropolitan Community. New York: Seminar Press. 1973.

[5] Erskine, Hazel, "The Polls: Women's Roles," Public Opinion Quarterly 35(Summer):275-290. 1971.

[6] Goldberg, Philip, "Are Women Prejudiced Against Women?" Trans-Action 5(April):28-30. 1968.

[7] Greeley, Andrew M. and Paul B. Sheatsley, "Attitudes toward Racial Integration," Scientific American 225(December):13-19. 1971.

[8] Lipman-Blumen, Jean, "How Ideology Shapes Women's Lives," Scientific American 226(January): 34-42. 1972.

[9] Mason, Karen Oppenheim and Larry L. Bumpass, "Women's Sex-Role Attitudes in the United States, 1970," Paper Presented at the Annual Meetings of the American Sociological Association, New York. August, 1973.

[10] Payne, Stanley L., The Art of Asking Questions. Princeton: Princeton University Press. 1951.

[11] Schwartz, Mildred A., Trends in White Attitudes toward Negroes. Chicago: National Opinion Research Center. 1967.

[12] Thiel, Henri, "On the Estimation of Relationships Involving Qualitative Variables," American Journal of Sociology 76(July):103-154. 1970.

[13] Vogel, Susan R., et alia, "Maternal Employment and Perception of Sex Roles Among College Students," Developmental Psychology 3(November): 384-391. 1970.

ACKNOWLEDGEMENT

This research was supported in part by a grant from the National Institute of Mental Health (Grant No. 1R01-MH25271-01).

ACCURACY OF 1970 CENSUS POPULATION AND HOUSING CHARACTERISTICS AS MEASURED BY REINTERVIEWS

Charles Jones, Henry Woltman, Kathryn Thomas and Stanley Cullimore
U.S. Bureau of the Census

Introduction

Reinterview as an evaluation method may be characterized in general terms as the selection of a sample of persons (or households), who, after answering the questions in the basic census or survey, are contacted at a later time and asked the same or similar questions again. The two responses for each person (or household) are compared to estimate the variability of responses in repeated interviews, and to estimate biases in the distribution of census or survey statistics. 1/

In the evaluation work of the Census Bureau, two general types of reinterview studies have been used. In the first type of study, each person (or household) is viewed as having an infinite set of responses to a specific question which can be generated by independent repetitions of the same survey procedure under the same general conditions. The initial census or survey obtains one of these responses while the reinterview obtains a second, by applying the same survey procedures under the same general conditions as existed in the initial interview. The two responses are assumed to have been randomly selected and are compared to produce estimates of the average trial-to-trial response variability, which is commonly referred to as simple response variance.

The second type of reinterview study is designed to obtain more accurate data than was feasible in the initial interview. These data are used as a standard of comparison for the initial census or survey responses. Here, the initial responses are viewed as being possibly defective because the enumerator may have been inadequately trained, the person answering the questions may not have been the most knowledgeable respondent, the questions and instructions may have been ambiguous, or other things of that nature. It is assumed that these deficiencies can be minimized in the reinterview by applying survey procedures such as the use of well-trained, highly qualified interviewers, choosing the most knowledgeable respondents to provide the data, applying detailed questioning sequences to probe those areas where the questions or instructions may have been ambiguous or inadequate, and reconciliation of differences in responses collected in the two interviews.

It is clear that neither of these two types of studies, in application, can meet their theoretical objectives. In both cases the estimates of response error have a tendency to be understated. In the first case, the conditions of the original interview cannot be duplicated in the reinterview to yield an independent response under the same general survey conditions. For example, the respondent, having answered the question once, is likely to be conditioned in his response in the second interview. The second type of reinterview study is unlikely to obtain the truth in all cases since the respondent may deliberately falsify his responses, or he may simply not know the answer to a particular question. Further, Census shares with all demographic survey organizations the problems of noninterviews. In our reinterviews we usually are not able to complete the study plans for all sample cases. For example, in carrying through the 1970 reinterview study we were unable to complete the study plan for about 25 percent of the persons and 20 percent of the housing units selected for reinterview.

Even with these limitations, however, reinterview is a valuable evaluation methodology. For example, following the 1950, the 1960, and the 1970 decennial censuses, reinterview studies were major components of the evaluation and research programs and these have produced useful data on response errors and their distributions.

Data were collected in the 1970 reinterview study for some 29 of the 100 plus questions included in the census. The questions selected were those for which the reinterview seemed to be an adequate vehicle for collecting reasonably accurate response error data, and for which another type of study would not produce the data more accurately or at less cost.

In the reinterview study, both of the reinterview techniques just described were used to evaluate the quality of the data collected in the 1970 census. For one set of characteristics, the reinterview questions approximated those used in the census. For another set of characteristics, including some of the new questions which lacked precise definitions on the specific subgroups of the population to be identified, more detailed questioning sequences were applied.

Comparison of the census and reinterview responses to a specific question for each person (or housing unit) yields the 2x2 table shown on the first page of the handout. In this table, the cells denoted as a and d represent sample counts where the census and reinterview responses agreed (i.e., either in or out of category) while the b and c cells represent sample counts where responses differed.

The effect of response errors on the quality of data collected for a particular category of a classification system is reflected by the levels of gross and of net error associated with that category. The gross error associated with a category represents the total number of response differences associated with the category (b+c), while the net error is the difference between the number erroneously included in the category and those erroneously omitted from the category (c-b). We have selected two summary response error measures to describe the level of gross and net error associated with the data collected in the census.

The summary measure of gross error - the index of inconsistency - is approximately the complement of the correlation between the census and reinterview responses. For example, if there is perfect positive correlation, the index is zero; if there is zero correlation, the index is 100. The index is

interpreted as the ratio of the observed response differences between two interviews to the response differences that would be expected if there was no correlation between interviews (i.e., if the two sets of responses were randomly associated). In terms of the handout table, the observed response differences are given by b and c, while the expected response differences would be estimated for those cells by the cross products of the marginal proportions in the two interviews. Under the conditions of the first type of reinterview discussed above, the index is also interpreted as the proportion of total variance accounted for by simple response variance. 2/ An index of inconsistency is estimated for each category of a distribution. In addition, a weighted average of the individual indices - referred to as the L-Fold index of inconsistency - provides a measure of the amount of inconsistency in the entire distribution. (When there are only two categories in the distribution, the L-Fold index is identical to the indices for the individual categories.) The range of the index (both for individual categories and for the entire distribution) is zero to 100. As a rule of thumb, we interpret an estimated index between 0 and 20 as indicating low inconsistency, between 20 and 50 moderate inconsistency, and above 50 high inconsistency.

The summary measure of net error - the net difference rate - provides a measure of bias in the census distribution when the reinterview responses are assumed to be more accurate than the census responses. This rate is simply the difference between the census and reinterview estimates of the proportion of housing units or persons in a given category of the distribution. A negative estimate indicates the census proportion is smaller than the reinterview; a positive estimate indicates the census proportion is larger than the reinterview proportion in the category.

Response Variance Reinterviews

For 11 population and housing characteristics the questions and procedures used in the reinterview were similar to those used in the census. That is, the methods used in the reinterview were not an attempt to collect an "improved" response, but simply to obtain responses to questions similar to those used in the census. 3/ The distributions for five of these 11 characteristics had, on the average, fairly low levels of response variance or inconsistency as the estimated L-fold indices were all under 20 (Table 1). The distribution for five of the characteristics had moderate levels of inconsistency, with L-fold indices between 24 and 45, while for one characteristic - Value of Home - the census and reinterview responses were highly inconsistent, as reflected by the estimated L-fold index of 58.

The estimated level of the index of inconsistency for a characteristic is sensitive to the detail of the classification system. Given the same set of responses, the estimate will usually increase as the detail increases and decrease as the detail decreases. For example, the index for Value of Home is 58 when the detail is the 11 categories in which the responses were collected on the census questionnaire. When the detail is reduced by forming fewer, broader categories, the index is also reduced. For example, looking at one of our published distributions for Value of Home which had only 6 categories, the index is estimated at 47. The indices shown in the handout were all estimated at the detail in which data were collected in the census. These would not apply to those published distributions where the detail is collapsed to broader categories. Additional evaluation of the Value of Home by use of a record check type study is being conducted. Those data will provide further insight into the accuracy of census responses for this question.

Response distributions for housing items were analyzed by occupancy status, tenure and size of structure. Generally, owners reported housing data more consistently than renters, responses for occupied units were more consistent than those for vacant units, and respondents in single unit structures reported more consistently than those in multi-unit structures.

Response error data for two of the 11 characteristics were available from the 1960 census - Bathtub or Shower Facilities and Flush Toilet Facilities. Comparison of 1970 and 1960 data for these items indicates that the response variability was of about the same order of magnitude in the two censuses (Table 3 - see notice on page 6 about table availability).

One question often asked is whether the use of a mail census increased or decreased the simple response variance in comparison with an enumerative canvass. Although the mail census technique is expected to have reduced the correlated component of response error contributed by enumerators, it is assumed that there may be differences in the simple response variance as well. A study included in the 1960 census estimated simple response variance for the two types of procedures in independent samples. The results indicated slight differences in simple response variances for the two types of procedures with the differences as observed tending to favor the mail procedure as having lower simple response variances. 4/ The 1970 reinterview study sheds little additional light on this question. Slight differences in response variance were observed for the two types of censuses, but these differences may reflect, in addition to procedural differences, differences in simple response variance for population groups covered by each type of census in 1970.

Response Bias Reinterviews

For 15 population and housing characteristics, a response bias type of reinterview was attempted. 5/ The reinterview involved the use of a detailed questioning sequence designed to probe areas where the question or instructions may have been unclear, and/or a reconciliation to obtain the "best" response when the census and reinterview classed the person or housing unit in different categories. The reinterviews were conducted in personal visits and the reinterviewer was instructed, when feasible, to observe the housing facilities, in addition to questioning the respondent. For population characteristics each adult member of the household responded for himself in the reinterview. The reinterviews were conducted by the permanent staff

of interviewers who work on the Bureau's Current Population Survey, after special training on reinterview techniques.

For 10 of these 15 characteristics the inconsistency of responses between the census and reinterview was low, the indices all falling below 20. The other five characteristics fell in the moderately inconsistent range with indices estimated between 21 and 45 (Table 2). The index for Number of Rooms appears to be particularly high relative to the other characteristics. Again, this partially reflects the detail in which these data were collected - one room intervals. We found that in about 80 percent of the cases with response differences, the census and reinterview differed by only one room. Collapsing this distribution to 5 categories, which was a published census distribution, the index is estimated at 38.

The consistency of reporting of housing characteristics by occupancy status, tenure, and size of structure generally followed the pattern discussed earlier; that is, owners were more consistent than renters, the reports for occupied units were more consistent than those for vacants, and data for single-units were more consistently reported than data for multi-unit structures. There was one notable exception - reports for Number of Rooms - where the owner/renter and single/multi-unit relationships were reversed. Considering the definitional problems associated with the number of rooms question, this reversal of relationship seems reasonable. Single family homes are more likely than multi-unit structures to contain space for which there may be confusion about whether it qualifies as a room according to the census definitions; for example, utility rooms or basements (finished or unfinished), enclosed porches, knee-walled attic space, etc.

There was corresponding response error data available from the 1960 census for four of these fifteen characteristics (Table 3). For Number of Children Ever Born, the L-Fold index associated with the 1960 census data was 10, indicating about the same level of inconsistency in both censuses. For Number of Bedrooms, the 1970 data had more consistency in reporting than in 1960. The 1970 L-Fold index was estimated at 18 and the 1960 L-Fold index was estimated at 26. Piped Water was also more consistently reported in 1970 than in 1960 with L-Fold indices respectively of 18 and 35. Number of Rooms was less consistently reported in 1970 than in 1960. The 1960 L-Fold index was estimated at 35 while in 1970 it was estimated at 44.

For five of these 15 characteristics moderate sized biases in census distribution were estimated. For Number of Children Ever Born, the "none" category was overstated in the census; that is, more women have had children than the census data indicate (Table 4). A review of the detailed reinterview responses, which we plan to undertake, may shed light on the source of this bias.

In Citizenship reports for the foreign born, the "naturalized" category was estimated as overstated in the census, while the "alien" category was estimated as understated. This seems to be the logical direction of the bias. Detailed review of the reinterview data will show whether persons undergoing the naturalization process, but not having completed it, were confused about the delineation between the two categories.

For Number of Bedrooms, there seemed to be some confusion by census respondents between the "none" and "one bedroom" categories; the "none" category was estimated to be understated in the census, the "one bedroom" category overstated. Here we think we know the source of the bias. For this characteristic, a review of the reinterview data indicates this confusion, for the most part, was associated with one-room efficiency apartments. The census definitions consider such units to have no bedroom. In a fair number of cases these units were reported by the census respondent as having one bedroom. It should be noted that as a part of the computer edits, a one room unit was edited as having no bedroom. Thus, the response errors which occurred in the field phase of enumeration probably were corrected as a result of that edit. We plan to follow through on this to learn how the processing edit may have affected the published distribution.

For Heating Equipment, the reinterview data indicate that biases exist in the census distribution for most of the heating equipment categories. There is some evidence that a fair proportion of the response errors are due to the respondents lack of knowledge, especially among householders in multi-unit structures. Errors in reporting heating equipment, as well as in other structural characteristics, may be reduced by collecting the data from a structure respondent (e.g. building manager, janitor, etc.). Data collected on the reinterview, but not yet analyzed, will indicate to what extent a structure respondent can improve response accuracy for these characteristics.

The response error data on Year Moved Into House reflects, in addition to respondent errors, a difference in the way the responses were recorded and edited in the two interviews. In the census, several year intervals were presented as possible answer categories with a final category of "always lived in this house or apartment." In the reinterview the specific year moved in was obtained and these were then coded to the appropriate category. In the edit of the reinterview responses, a child born after his parents had moved into a house had his response edited to the "always lived in the same house" category. In the census many of these children were reported as moving into the house in their birth year, rather than being reported in the "always lived in same house" category. The census responses for these children were not edited to the "always lived in same house" category during processing. The reinterview estimate of the percent classified in the "always lived in same house" category is on the order of 7 to 9 percentage points higher than the corresponding census percent (Table 5). Most of the understatement in the census is concentrated in the son or daughter of head population group. These data suggest that the consistency check, if applied to census responses, would appreciably improve the accuracy of the published census distribution.

Reinterview For New Population Questions

A third type of reinterview analysis was made a part of the 1970 reinterview study. Data for Mother Tongue (for the total population), Spanish Origin or Descent, and Vocational Training, were obtained for the first time in the 1970 Census. These questions or concepts were not precisely defined in the census and, as a result, the questions were subject to broad ranges of interpretation. For these characteristics, the development of a "correct" response in the reinterview did not appear to be fruitful. Thus, the reinterview focused on collecting detailed data to provide insights into how respondents interpreted the census questions.

Mother Tongue

The census question from which data on Mother Tongue were derived inquired as to "What language other than English was spoken in the person's home when he was a child?" Five answer categories of "Spanish", "French," "German," "Other foreign" (with a write in entry for the specific language) and "English only," were provided in the body of the question. In the reinterview, each person was questioned in detail to determine if any foreign language was used in his childhood home. For persons reporting use of a foreign language in the childhood home, the reinterview obtained data on the frequency of foreign language usage relative to English and on who spoke the language in the childhood home (i.e., person himself, parents, grandparents, etc.).

These reinterview data were used to stratify persons reporting a foreign language usage in their childhood home by the degree or intensity of usage. Table 6 provides the definition of each of the strata or levels and shows the distribution of census responses within each level. Level 1, for example, represents those persons for whom the foreign language was the only language used in the childhood home; English was reportedly not used in the childhood home. At the other extreme, level 7, the foreign language was not spoken by the sample person himself, but was spoken by other family members. Moving from level 1 to level 7, the reporting completeness in the census drops off from 97 percent to about 52 percent (Table 6, column 5).

Summary measures of response error, which result when alternative definitions of the population to be identified by the question are considered, are provided in Table 7. Using the broadest definition of Mother Tongue, levels 1 through 7, the proportion of persons reporting the use of a foreign language in the childhood home in the census is understated between 4 and 5 percentage points. The index of inconsistency for this reinterview definition is the lowest among the alternatives considered, estimated at 18. As the definition becomes more restrictive, the proportion of persons reporting use of a foreign language in their childhood home in the census exceeds the corresponding reinterview proportion. The index of inconsistency also increases as the definition becomes more restrictive. These data indicate that respondents tended to apply a literal interpretation to the census question, reporting any foreign language usage in the childhood home regardless of the intensity of usage.

Both errors in reporting use of a foreign language in the childhood home as well as errors in reporting the specific foreign language spoken affects the Mother Tongue distribution of languages. Table 8 shows that when both sources of error are considered, the census reports for 8 language categories were fairly consistent with those derived from the reinterview data, using levels 1 through 7 to identify persons having a foreign language spoken in their childhood home. The indices of inconsistency for these categories range from 11 to 24. The majority of the inconsistencies result from differences in reporting use of a foreign language in the childhood home, and, to a much lesser extent, the inconsistencies in reporting the specific foreign language spoken. Evidence of this is provided in Table 9 which shows the indices of inconsistency for the language categories which reflect only differences in reporting the specific foreign language spoken (i.e., these indices are based on persons reporting use of a foreign language in both the census and reinterview). These indices range from 1 to 8.

Spanish Origin or Descent

The census question on Spanish Origin or Descent asked if the person's origin or descent was "Mexican," "Puerto Rican", "Cuban," "Central or South American," "Other Spanish," or "none of these". The reinterview probed in detail about Spanish ancestry on either side of the family, and if such ancestry was detected, questioned about who the ancestors were (e.g., father, grandmother, great grandparent) and the country from which they came.

The census reports for persons of Spanish Origin or Descent were moderately inconsistent with those obtained from the detailed reinterview data; the index of inconsistency is estimated at 22 (Table 10). The proportion of persons reporting Spanish Origin or Descent in the census is slightly lower than the level estimated in the reinterview (-0.3 percent).

Analysis of the data for selected subgroups of the population indicate that major differences in the consistency of reporting exist, as might be expected, by Spanish surname, by nativity, by race, and by major geography.

A review of responses for persons of Spanish origin in the reinterview by their census responses indicates that reporting of Spanish origin in the census was correlated with (a) whether the origin was on both sides of the family or only on one side of the family, (b) which ancestors were from a Spanish speaking country, and (c) the country of Spanish origin (Table 11).

For example, of the persons with Spanish origin on both sides of the family according to the reinterview, about 97 percent reported themselves as of Spanish origin in the census, while only 21

percent of the persons with Spanish origin on only one side of the family reported themselves as of Spanish origin in the census. When the sample person himself was from a Spanish speaking country, Spanish Origin or Descent was almost always reported in the census (estimated at 97 percent). When the Spanish ancestry was a parent or grandparent, the reporting of Spanish origin in the census was about 80 percent (estimates of 83 and 73 percent). Less than 50 percent reported Spanish origin in the census when the ancestry was further back than grandparent. Reporting of Spanish origin in the census for persons with Spanish origin from Mexico, Cuba, Puerto Rico, or a Central or South American country was estimated at about 90 percent. On the other hand, only about 30 percent of the persons with Spanish origin from some other Spanish speaking country (mostly Spain) reported Spanish origin in the census.

We also observed that among persons not of Spanish origin according to the reinterview, a small number were reported as of Spanish origin in the census. In a third interview (reconciliation) with these persons we learned that these were primarily persons born in southern or midwestern states of the U.S. who had misinterpreted the "Central or South American" response category.

Vocational Training

The census question on vocational training had two parts. The first part asked the respondent, "Has this person ever completed a vocational training program?", while the second part asked the main field of vocational training for those persons reporting completion of a vocational training program. Some examples of programs which were to be reported as vocational were included in the body of the question and additional instructions were provided the respondents and enumerators as to types of programs which were not to be reported.

In the reinterview a detailed battery of questions was used to identify any training experience that might be considered as vocational. These detailed data also provided a basis for identifying persons having training experiences which were clearly not to be reported as vocational according to the census instructions.

The comparison of the reinterview and census responses shows a large number of response differences associated with the question. Of persons who had completed a vocational training program according to the reinterview (Table 12; 1415 cases), about 39 percent did not report completing a vocational training program on the census questionnaire (Table 12; 556 cases). Conversely, of persons who did not complete a vocational training program according to the reinterview (Table 12; 6152 cases), some 8½ percent reported completing a program in the census (Table 12; 524 cases). These relatively large numbers of response differences tended to offset each other so that, on a net basis, the proportion of persons reporting vocational training was estimated to be approximately the same in the reinterview and census. However, the consistency in reporting vocational training between the census and reinterview was quite low as evidenced by the estimated index of inconsistency of 47.

We expected that reporting completeness in the census would be higher for some types of training programs than others - for example, lengthy programs, training actually used on a job, etc. These expectations, for the most part, were not realized. The reinterview data (Table 13) indicate that the proportion of completed training programs classified as vocational in the reinterview which were reported as completed in the census, was not highly correlated with the field of training, the year the program was completed, where the training was received (i.e., trade or technical school, other type of school, not in school) man hours spent in the program, or the usefulness of the training (i.e., currently being used on the job, previously used on the job, or never used). In general, each of these categories was subject to substantial incompleteness of reporting in the census, although minor differences in reporting completeness were observed between some categories.

The reasons for persons erroneously reporting completion of a vocational training program in the census were fairly diverse but they seem to be related to a failure to follow instructions. The data given in Table 14 indicate that the major source of these errors included those resulting from persons reporting completion of a vocational training program when in fact, they had attended a program but had not completed it, and the reporting of academic training, on-the-job training, and training taken in a company school. These were types of training which the respondent and enumerators were specifically instructed not to report in the census.

For persons reporting completion of a vocational training program in both the census and reinterview, the census responses to the Field of Training questions were highly consistent with those obtained in the reinterview. The L-Fold index of inconsistency for the main field of training distribution was estimated at 9 (Table 15.)

1/ The data presented in this report describe the quality of responses recorded on the census questionnaires at the field stage of enumeration. They do not reflect the effect of errors, corrections or additions made during clerical and computer processing, in preparing the data for publication.

2/ The simple response variance is the between trial variation in responses averaged over all persons. Another component of response variance reflects the correlation of response deviations within trial. This component may be introduced by the enumerator, coder, crew leader, etc. The correlated component due to enumerators is being estimated in another study in the 1970 Evaluation Program.

3/ There were, naturally, some differences between the two surveys that may have affected the expected values of the responses, such as the use of well-trained permanent staff of

current survey interviewers to collect the reinterview data rather than temporary enumerators who collected the census data. On balance, however, we believe the reinterview more closely approximates a response variance type reinterview than a response bias type reinterview.

4/ U.S. Bureau of the Census. Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Effects of Different Reinterview Techniques on Estimates of Simple Response Variance, Series ER60, No. 11 U.S. Government Printing Office, Washington, D.C. 1972.

5/ In this type of reinterview, the index of inconsistency is best interpreted as approximately the complement of the correlation between the census and reinterview responses. The alternate interpretation, as the ratio of simple response variance to the total variance, is of questionable validity as the reinterview techniques used tend to introduce a downward bias in the estimate of simple response variance.

NOTICE

Due to space limitations, all the detailed tables on the handout cannot be reproduced here. Those tables essential for understanding the discussion in the text are reproduced below. A complete copy of the handout may be obtained by writing any of the authors at the following address: Statistical Methods Division, Bureau of the Census, Washington, D.C. 20233

HANDOUT

Analysis of response accuracy for a given characteristic between two trials is illustrated as follows: Each element (person or housing unit) is treated as distributed (0,1) in each trial. Responses for a given element in the two trials are compared and the element is placed in one of the internal cells of the 2X2 table. Comparison of responses for each element, over a sample of n elements, generates the entire 2X2 table. (For a characteristic with more than two categories the distribution is collapsed into a series of 2X2 tables, one for each category).

Reinterview Response (TRIAL 2)	Census Response (TRIAL 1)		
	In Category (1)	Not in Category (0)	Total
In Category (1)	a	b	np_2
Not in Category (0)	c	d	nq_2
Total	np_1	nq_1	n

The attached tables present summary measures of response errors for Population and Housing Characteristics as estimated from the reinterview studies. Two summary measures are presented. One describes the gross error while the second describes the net error in the distributions between the two surveys.

The measure of gross error - The Index of Inconsistency - is approximately the complement of the correlation between the census and reinterview responses. It is interpreted as the proportion of total population variance accounted for by simple response variance or as the ratio of the observed response differences between the two trials to the response differences that would be expected if there was no correlation between the two trials (i.e., if the two sets of responses were randomly associated). The index ranges from 0 to 100 and as a rule of thumb an estimated index between 0 and 20 indicates low inconsistency, between 20 and 50 moderate inconsistency, and above 50 high inconsistency. An index is estimated for each category of a distribution.

In the notation of the above table the index is estimated as: $\hat{I} = \left[\frac{\frac{b+c}{n}}{p_1q_2+p_2q_1}\right] X100$

In addition, a weighted average of the individual indices is estimated - referred to as the L-Fold Index of Inconsistency - which describes the amount of inconsistency in the entire distribution, and is interpreted in the same way as the indices for individual categories. (When there are only two categories in the distribution the L-Fold index is identical to the indices for the individual categories.)

In the notation given above the L-Fold index is estimated as: $\hat{I}_L = \sum_{i=1}^{L} \hat{I}_i \cdot \left[\frac{(p_{i1}q_{i2} + p_{i2}q_{i1})}{\sum_{i=1}^{L} (p_{i1}q_{i2} + p_{i2}q_{i1})}\right]$

The summary measure of net error - The Net Difference Rate - provides an estimate of the bias in the census distribution when the reinterview responses are assumed to be more accurate than the census responses. This rate is simply the difference between the reinterview and census estimates of the proportion of persons or housing units in a given category of the distribution.

In the notation of the above table the net difference rate is estimated as: $NDR = (p_1-p_2) X100$

TABLE 1 - L-Fold Index of Inconsistency for Population And Housing Characteristics Estimated From a Response Variance Type Reinterview, 1970 Census

(All estimates have been multiplied by 100.)

Characteristic (No. of Categories in Distribution)	L-Fold Index of Inconsistency @ (1)	95-Percent Confidence Interval for L-Fold Index (2)
Heating Fuel, Occupied Units (7)	12	10.2 to 13.3
Owner Occupied	8	6.8 to 10.0
Renter Occupied	19	15.8 to 22.4
Units in Single Unit Structures	9	7.5 to 10.7
Units in Multi-Unit Structures	20	16.6 to 24.6
Renters Paying Extra for Electricity, All Units (2)	15	11.7 to 18.5
Renters Paying Extra for Gas, All Units (2)	18	15.1 to 21.9
Bathtub or Shower Facilities, All Units (3)	18	15.3 to 21.3
Occupied Units	17	14.3 to 20.7
Vacant Units	24	16.9 to 34.2
Units in Single Unit Structures	15	11.8 to 18.1
Units in Multi-Unit Structures	29	21.4 to 39.1
Flush Toilet Facilities, All Units (3)	18	15.4 to 21.7
Occupied Units	16	13.4 to 20.0
Vacant Units	30	21.2 to 41.1
Units in Single Unit Structures	14	11.4 to 18.0
Units in Multi-Unit Structures	31	22.3 to 42.2
Telephone Availability, Occupied Units (2)	24	21.1 to 26.6
Owner Occupied	22	18.8 to 26.8
Renter Occupied	26	22.5 to 30.6
Year Structure Built, All Units (6)	25	24.0 to 26.5
Owner Occupied	22	20.2 to 23.0
Renter Occupied	36	33.0 to 38.6
Occupied Units in Single Unit Structures	25	23.1 to 26.0
Occupied Units in Multi-Unit Structures	29	25.8 to 31.9
Vacant Units	50	40.8 to 63.1
Vacancy Status (2)	31	24.4 to 39.3
Renters Paying Extra for Water, All Units (2)	39	31.4 to 49.7
Renters Paying Extra for Other Fuels, All Units (2)	45	34.1 to 59.3
Value of Home, Occupied Units (11)	58	56.5 to 60.2

@ The level of the L-fold index is sensitive to the detail of the classification system. For example if Value of Home data were collected in $100 class intervals we would expect to observe many more response diffences between trials, and to obtain a higher estimated L-fold index, than if the data were collected in $10,000 class intervals. The indices shown here were estimated at the detail to which data were collected in the census. These indices would not apply to published distributions where the detail data, as collected, were collapsed to broader categories.

TABLE 2 - L-Fold Index of Inconsistency for Population And Housing Characteristics Estimated From a Response Bias Type of Reinterview, 1970 Census

(All estimates have been multiplied by 100.)

Characteristic (No. of Categories in Distribution)	L-Fold Index of Inconsistency @ (1)	95-Percent Confidence Interval for L-Fold Index (2)
Nativity (2)	1	0.8 to 2.2
Nativity-Father(2)	3	2.1 to 3.4
Nativity-Mother(2)	3	2.1 to 3.4
Tenure, Occupied Units(4)	4	3.6 to 5.1
Contract Rent, Occupied Units(14)	11	9.5 to 12.8
Units in Single Unit Structures	9	6.8 to 11.7
Units in Multi-Unit Structures	12	10.5 to 14.6
Citizenship; Foreign Born (3)	12	8.6 to 17.2
Number of Children Ever Born*(13)	12	11.2 to 13.2
Year Came to U.S. To Stay, Foreign Born(9)	13	10.4 to 17.4
Number of Units in Structure, All Units (10)	15	13.5 to 15.9
Number of Bedrooms, All Units (6)	18	17.0 to 20.1
Occupied Units	18	16.7 to 19.8
Vacant Units	33	23.1 to 48.8
Piped Water, All Units (3)	21	18.2 to 25.0
Occupied Units	18	14.7 to 21.5
Vacant Units	41	31.7 to 53.5
Units in Single Unit Structures	20	17.0 to 24.1
Units in Multi-Unit Structures	41	23.6 to 70.7
Year Moved Into House (6)	25	24.6 to 26.5
Kitchen Facilities, Occupied Units (3)	25	20.4 to 30.6
Units in Single Unit Structures	21	16.7 to 26.8
Units in Multi-Unit Structures	46	30.6 to 68.9
Heating Equipment, All Units (8)	27	26.0 to 28.5
Occupied Units	27	25.3 to 27.9
Vacant Units	53	45.0 to 64.7
Units in Single Unit Structures	25	23.8 to 26.7
Units in Multi-Unit Structures	35	32.0 to 37.7
Owner Occupied	25	23.1 to 26.3
Renter Occupied	32	29.6 to 34.1
Number of Rooms, All Units (9)	45	44.1 to 46.8
Owner Occupied	51	49.6 to 53.0
Renter Occupied	37	35.0 to 39.6
Units in Single Unit Structures	50	48.6 to 51.9
Units in Multi-Unit Structures	34	31.7 to 37.1
Occupied Units	45	43.9 to 46.6
Vacant Units	55	47.4 to 65.0

* Based on reports for ever married females 14 to 64 years old in both Census and Reinterview.

@ The level of the L-fold index is sensitive to the detail of the classification system. For example if Value of Home data were collected in $100 class intervals we would expect to observe many more response differences between trials, and to obtain a higher estimated L-fold index, than if the data were collected in $10,000 class intervals. The indices shown here were estimated at the detail to which data were collected in the census. These indices would not apply to published distributions where the detail data, as collected, were collapsed to broader categories.

TABLE 6 - Comparison of Census and Reinterview Responses For Foreign Language Usage in Childhood Home, 1970 Census

(Data shown as numbers of sample persons reinterviewed and matched to census questionnaires, not as inflated estimates.)

Reinterview Classification	Census Response					
	Number			Percent Distribution		
	Total Persons (1)	Foreign Language Spoken in Childhood Home (2)	English Only Spoken in Childhood home (3)	Total Persons (4)	Foreign Language Spoken in Childhood Home (5)	English Only Spoken in Childhood home (6)
Total Persons	11,102	1,655	9,447	100.0	14.9	85.1
Foreign Language Spoken	2,170	1,627	543	100.0	75.0	25.0
Spoken by Person Himself	1,383	1,216	167	100.0	87.9	12.1
Foreign Language Only Spoken (Level 1)	412	399	13	100.0	96.8	3.2
Foreign Language Predominant, English Also Spoken (Level 2)	445	410	35	100.0	92.1	7.9
Foreign Language Spoken Equally With English (Level 3)	41	38	3	100.0	92.7	7.3
English Predominant, Foreign Language Spoken Frequently 1/ (Level 4)	350	276	74	100.0	78.9	21.1
English Predominant, Foreign Language Spoken Occasionally 2/ (Level 5)	95	72	23	100.0	75.8	24.2
English Predominant, Foreign Language Spoken Seldom 3/ (Level 6)	40	21	19	100.0	52.5	47.5
Not Spoken by Sample Person But Spoken by Other Family Members (Level 7)	787	411	376	100.0	52.2	47.8
English Only Spoken	8,932	28	8,904	100.0	0.3	99.7

1/ For example, spoken daily in the home.

2/ For example, spoken when relatives visited or to keep outsiders from understanding conversation.

3/ For example, used for slang, phrases, expressions.

TABLE 10 - Summary Measures of Response Error for Reporting Spanish Origin or Descent, by Selected Characteristics, 1970 Census

(All estimates have been multiplied by 100)

Characteristic	Index of Inconsistency (1)	95 Percent Confidence Interval for Index of Inconsistency (2)	Percent in Class in Reinterview (3)	Net Difference Rate @ (4)	95 Percent Confidence Interval for Net Difference Rate (5)
Total	22	18.6 to 25.7	4.0	-0.3	-0.6 to -0.1
Spanish Surname	11	7.3 to 17.0	56.9	-3.5	-5.9 to -1.1
No Spanish Surname	49	41.1 to 58.5	1.6	-0.2	-0.4 to -0.1
Native born	27	22.9 to 32.1	3.3	-0.4	-0.6 to -0.1
Foreign born	5	2.5 to 10.9	17.2	0.2*	-1.0 to 1.4
Age 0 to 19	25	20.0 to 31.9	5.1	-1.0	-1.5 to -0.4
Age 20 to 44	17	12.4 to 23.2	4.5	0.0*	-0.5 to 0.4
Age 45 or older	23	16.8 to 32.8	2.4	0.1*	-0.3 to 0.5
Male	21	16.8 to 26.8	4.2	-0.3*	-0.7 to 0.1
Female	22	17.8 to 28.2	3.8	-0.4*	-0.7 to 0.0
Son or Daughter of Head	25	19.6 to 32.2	4.8	-1.0	-1.6 to -0.5
Not Son or Daughter of Head	20	16.2 to 24.9	3.6	0.1*	-0.3 to 0.4
White	20	16.6 to 24.0	4.0	-0.4	-0.7 to -0.1
Negro	88	59.7 to 100.0	1.4	0.7*	-0.5 to 1.9
Other Races	11	4.2 to 28.4	17.9	-3.0*	-7.8 to 0.4
1/ Southwest	11	8.4 to 15.6	14.6	-2.0	-2.9 to -1.2
East	28	21.4 to 37.5	2.8	-0.1*	-0.5 to 0.3
Midwest	35	22.0 to 56.0	1.0	0.9	0.4 to 1.4
Balance of U.S.	69	50.3 to 95.2	1.4	-0.4*	-1.0 to 0.1
Conventional census areas	28	21.1 to 37.7	2.6	-0.1*	-0.5 to 0.3
Mail census areas	20	16.3 to 24.2	4.7	-0.4	-0.8 to -0.1

* Indicates net difference rate is not significantly different from zero at the 95 percent confidence level.

@ Difference between census and reinterview estimates of percent in class. Netative estimate indicates census less than reinterview; positive estimate indicates census larger than reinterview.

1/ Southwest: Arizona, California, Colorado, New Mexico, Texas
East: Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware, Maryland, Washington,D.C., Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida
Midwest: Illinois, Indiana, Ohio, Michigan
Balance of U.S.: All other states not mentioned above.

TABLE 19 - Evaluation of Census Report By Type of Vocational Training Program Completed, 1970 Census

(Data shown as sample counts of persons 14 years and older who, according to the reinterview had completed any training program which might be considered vocational, other than those which were not to be reported according to the census instructions.)

Category	Total Persons	Reported as Completed in the Census		95 Percent Confidence Interval For Percent
		Number	Percent	
Persons classified as completing vocational training program on the basis of the reinterview, total.....	1,415	859	61	58.4 to 63.6
Field of Training				
Business.....	515	305	59	54.6 to 63.4
Nursing.....	183	114	62	55.0 to 69.0
Trades.....	551	339	62	57.8 to 66.2
Engineer.....	68	42	62	50.2 to 73.8
Agriculture.....	43	25	58	43.0 to 73.0
Other.....	15	11	73	50.2 to 95.8
Field not reported	37	-	-	-
Year Program Completed				
1969 or later.....	89	52	58	47.6 to 68.4
1965-1968.....	252	139	55	48.8 to 61.2
1960-1964.....	170	111	65	57.8 to 72.2
1950-1959.....	295	185	63	57.4 to 68.6
1940-1949.....	307	198	64	58.6 to 69.4
1939 or earlier.....	233	141	60	53.6 to 66.4
Year not reported.....	69	-	-	-
Where or How Training Received				
Trade or technical school	636	425	67	63.2 to 70.8
Other school or not in school	682	372	54	50.2 to 57.8
High School.....	249	114	46	39.6 to 52.4
College.....	147	79	54	45.8 to 62.2
Other School.....	127	81	64	55.6 to 72.4
Not in school 1/....	159	98	62	54.2 to 69.8
Where or how received not reported	97	-	-	-
Man Hours in Program				
Less than 100 Man Hours	188	80	42	34.8 to 49.2
Under 25 Man-Hours	40	12	30	15.6 to 44.4
25-99 Man-Hours....	148	68	46	37.8 to 54.2
100 or More Man Hours...	1,119	715	64	61.2 to 66.8
100-249 Man-Hours..	176	86	49	41.6 to 56.4
250-499 Man-Hours..	137	73	53	44.6 to 61.4
500-999 Man-Hours..	209	134	64	57.4 to 70.6
1,000-1,999 Man-Hours	312	210	67	61.8 to 72.2
2,000 or More Man-Hours	285	212	74	69.2 to 79.6
Man Hours Not Reported	108	-	-	-
Usefulness of Training 2/				
Used in Current Job.....	618	410	66	62.2 to 69.8
Previously Used or Never Used in Job	779	438	56	52.4 to 59.6
Previously Used.....	504	295	58	53.6 to 62.4
Never Used, training sufficient to qualify for civilian job in that field.	223	121	54	47.4 to 60.6
Never Used, training not sufficient to qualify for civilian job in that field.	52	22	42	28.4 to 55.6
Usefulness not reported	18	-	-	-

1/ Includes training received through the Job Corps, or an apprenticeship, and military training which could be used in a civilian job.

2/ Respondent's assessment

ACCURACY OF CENSUS DATA AS MEASURED BY THE 1970 CPS-CENSUS-IRS MATCHING STUDY

Paula Schneider and Joseph Knott
Social and Economic Statistics Administration, Bureau of the Census

INTRODUCTION

Among the various projects designed to evaluate the quality of statistics from the 1970 Census of Population is the CPS-Census-IRS Matching Study. The data in this study are based on case-by-case comparisons of persons enumerated in the March 1970 Current Population Survey and in the census 20-percent sample with regard to classification by a variety of demographic, social, and economic characteristics. In addition, for a portion of the universe the census income data were compared with information as recorded on 1969 Federal income tax returns. This paper deals briefly with the methods and design of the study and presents a capsulized view of the quality of various characteristics, as measured by the index of inconsistency and the net difference rate. All figures and statements contained herein are, however, preliminary and subject to revision. Final detailed statistics on these subjects will be published in the report series covering the Evaluation and Research Program of the 1970 Census of Population and Housing [PHC(E)].

The universe for this study was restricted to persons enumerated as members of households in the March 1970 CPS and in decennial census 20-percent households. To accomplish this match of identical persons from the two sources, the following operations were undertaken. A 1970 census geographic identification (i.e. enumeration district) was assigned to each housing unit in the March 1970 CPS and a search was made for the unit in the appropriate census address register. If an address match was made to a census unit listed as having been enumerated on a 100-percent or "short form" questionnaire, the CPS unit was dropped from the scope of this study although it was included for an associated study designed to measure population coverage in the 1970 census. If the CPS address was matched to a census 20-percent sample household, the census questionnaire was obtained and a name match performed. The matched cases constituted the universe for this study.

For each matched unit, selected CPS and census identification information on the household and persons therein was transcribed to a specially designed, machine-readable control sheet and then transferred to computer tape. This control tape was matched against the March 1970 CPS data file and to the census sample data file from which records for the appropriate persons were obtained.

Since the information was obtained from the final edited CPS and census data files, comparisons between the data sources can be used to estimate errors in publication level statistics. However, since the data are, by definition, restricted to persons for whom records were located in both sources, the measures of error reflect only content differences and not any error due to coverage problems.

In determining the levels of response error in census statistics by comparing the census classification with the corresponding classification in the Current Population Survey (CPS), the CPS response is assumed to be more accurate for some characteristics. This assumption is made since the CPS is a monthly national survey which utilizes a staff of full-time, experienced interviewers and is conducted under more extensive controls and training procedures. However, there are certain limitations involved in estimating response error by this method. First, in such a study it is seldom possible to locate the data records for all persons in both sources. In the 1970 CPS-Census Match, we were able to obtain matched records for about 75 percent of persons in the sample. If the response error distributions of the unmatched cases were generally different from those for the matched population, the data would be biased. For the purpose of the analysis in this report the assumption was made that they did not differ to an appreciable degree. Second, even though the CPS response is usually assumed to be the standard of accuracy, the CPS is obviously subject to some degree of error. In fact, for some characteristics, such as age, the CPS may be as error prone as the census. Third, whereas the CPS data are obtained through personal interview, the census data are based partially on self-enumeration responses and partially on personal interview. Therefore, differences in the type of enumeration and in the household member(s) being interviewed or completing the questionnaire may have had some effect on the responses given. The final factor to be considered in interpreting differences between the CPS and census data is the variation in time of enumeration. The census period generally extended from the last week in March and several weeks during April 1970, or longer in some areas, as compared with the 1-week enumeration period (March 16 to 20) for CPS. Therefore, some of the differences observed are correct in the sense that changes in status occur over time.

MEASURES OF RESPONSE ERROR

Two measures of response error have been estimated for each subject characteristic. These measures are the index of inconsistency - a measure of gross error or response variability - and the net difference rate - a measure of net error or bias. Estimated values of the index of inconsistency have been computed for each category in a distribution and a weighted average of the individual indices, the L-fold index of inconsistency, provides an estimate of the overall consistency of classification between the CPS and census for a given characteristic. Also, in table 1 the number of categories in the distribution from which the L-fold index was derived is shown in parentheses following each characteristic. For the purpose of measuring the adequacy of the census data collection system for a particular characteristic, an index under 20 can be considered to indicate a relatively low level of inconsistency, those between 20 and 50 a moderate level, and those over 50 a high level of inconsistency. In terms of published cross-tabulations of census data, the index

Table 1.—L-FOLD INDEX OF INCONSISTENCY FOR SELECTED POPULATION CHARACTERISTICS FROM THE C-S-CENSUS MATCHING STUDY: 1970 AND 1960

Characteristic (number of categories in distribution)	1970 L-fold index of inconsistency 1/	1970 95% confidence interval on index of inconsistency	1960 L-fold index of inconsistency 1/
AGE (17)			
Total	7	6.8 to 7.5	6
Male	7	6.1 to 7.1	5
Female	8	7.0 to 8.1	6
White	7	6.2 to 7.0	5
Negro 2/	12	10.9 to 14.1	11
Metropolitan-central city	9	7.9 to 9.5	(NA)
Metropolitan-outside central city	6	5.6 to 6.8	(NA)
Nonmetropolitan-urban	6	5.2 to 7.1	(NA)
Nonmetropolitan-rural	7	6.6 to 8.4	(NA)
SEX 3/ (2)	2	1.8 to 2.4	3
RACE (3)	3	2.8 to 4.0	4
HOUSEHOLD RELATIONSHIP			
Male (5)	4	3.6 to 4.6	4
Female (6)	5	4.1 to 5.1	5
MARITAL STATUS (5)			
Male, 14 years and over	5	4.0 to 5.5	6
Female, 14 years and over	5	4.3 to 5.6	5
EDUCATIONAL ATTAINMENT (13)			
Male, 14 years and over	38	37.2 to 39.8	(NA)
Female, 14 years and over	37	35.9 to 38.3	(NA)
VETERAN STATUS AND PERIOD OF SERVICE 4/ (6)	15	13.6 to 16.3	(NA)
EMPLOYMENT STATUS 5/ (4)			
Male	19	17.9 to 20.9	19
Female	20	19.1 to 21.7	20
WORK EXPERIENCE 6/ (7)			
Male	43	41.6 to 45.2	43
Female	37	35.9 to 38.8	36

Characteristic (number of categories in distribution)	1970 L-fold index of inconsistency 1/	1970 95% confidence interval on index of inconsistency	1960 L-fold index of inconsistency 1/
CLASS OF WORKER 7/ (7)			
Male	16	14.3 to 17.7	(NA)
Female	18	15.7 to 20.6	(NA)
MAJOR OCCUPATION 7/ (12)			
Male	31	29.4 to 32.3	25
Female	21	19.3 to 22.8	15
MAJOR INDUSTRY 7/ (12)			
Male	17	16.2 to 18.6	15
Female	14	12.3 to 15.2	10
PERSONS TOTAL MONEY INCOME (15)			
Total	45	44.0 to 45.7	38
Male	50	48.6 to 51.1	46
Female	43	41.9 to 44.5	33
WAGE OR SALARY INCOME (14)			
Male	39	38.0 to 40.5	(NA)
Female	33	31.5 to 34.2	(NA)
NONFARM SELF-EMPLOYMENT INCOME (15)			
Male	57	52.8 to 61.9	(NA)
Female	66	57.7 to 75.0	(NA)
FARM SELF-EMPLOYMENT INCOME (15)			
Male	48	42.5 to 53.4	(NA)
Female	91	68.4 to 100.0	(NA)
SOCIAL SECURITY INCOME (7)			
Male	24	21.9 to 26.8	(NA)
Female	30	27.5 to 32.3	(NA)
PUBLIC ASSISTANCE INCOME (7)			
Male	52	44.2 to 61.7	(NA)
Female	45	39.7 to 51.3	(NA)
ALL OTHER INCOME (15)			
Male	59	57.2 to 61.4	(NA)
Female	57	54.4 to 59.8	(NA)

NA Not available.

1/ The level of the L-fold index is sensitive to the detail of the classification system. For example, if age data were classified in one-year intervals, we would expect to observe more differences between trials and to obtain a higher estimated L-fold index than if the data were classified in five-year intervals. The indices shown here would not apply to published distributions where the data were either shown in more detailed or in broader categories.

2/ Refers to "Negro and other races" for 1960.

3/ Since there are only two categories in the distribution, the index of inconsistency is not an average measure.

4/ Based on civilian males 16 years old and over.

5/ Based on civilian population 14 years old and over.

6/ Based on civilian population 16 years old and over.

7/ Based on persons 16 years old and over, employed in CPS and census.

Table 2.—INDEX OF INCONSISTENCY AND NET DIFFERENCE RATE FOR SELECTED POPULATION CHARACTERISTICS FROM THE 1970 CPS-CENSUS MATCHING STUDY

Characteristic	Index of inconsistency	95 percent confidence interval on index of inconsistency	Percent in class (CPS)	Net difference rate[1/]	95 percent confidence interval on net difference rate
AGE					
Under 1 year	6	4.7 to 7.3	3.2	0.0*	-0.1 to 0.0
1 to 4 years	7	6.2 to 8.5	5.3	-0.1*	-0.2 to 0.1
5 to 9 years	5	3.9 to 5.3	10.5	0.0*	-0.1 to 0.2
10 to 14 years	4	3.5 to 4.7	11.1	-0.1*	-0.2 to 0.0
15 to 19 years	5	4.0 to 5.4	8.9	0.0*	-0.1 to 0.1
20 to 24 years	6	5.4 to 7.4	6.2	0.1*	-0.1 to 0.2
25 to 29 years	7	5.6 to 7.6	6.5	0.0*	-0.1 to 0.1
30 to 34 years	9	8.2 to 10.7	5.7	0.0*	-0.1 to 0.2
35 to 39 years	9	7.7 to 10.2	5.9	-0.1*	-0.2 to 0.1
40 to 44 years	9	7.6 to 9.9	6.1	0.0*	-0.2 to 0.1
45 to 49 years	8	7.0 to 9.3	6.2	0.0*	-0.1 to 0.2
50 to 54 years	8	7.3 to 9.7	5.8	0.0*	-0.1 to 0.1
55 to 59 years	9	7.6 to 10.2	5.1	0.0*	-0.1 to 0.1
60 to 64 years	10	8.9 to 12.0	4.1	0.0*	-0.1 to 0.1
65 to 69 years	12	9.9 to 13.5	3.5	-0.1*	-0.2 to 0.1
70 to 74 years	10	8.5 to 12.3	2.8	0.1*	0.0 to 0.2
75 years and over	8	6.5 to 9.6	3.3	0.1*	0.0 to 0.2
SEX					
Male	2	1.8 to 2.4	48.0	0.0*	-0.2 to 0.1
Female	2	1.8 to 2.4	52.0	0.0*	-0.1 to 0.2
RACE					
White	3	2.8 to 4.0	90.5	-0.3	-0.4 to -0.2
Negro	1	1.0 to 1.8	8.7	0.1*	0.0 to 0.1
Other races	19	15.5 to 24.2	0.9	0.2	0.1 to 0.3
HOUSEHOLD RELATIONSHIP					
MALE					
(Primary) family head	2	1.7 to 2.5	49.5	0.0*	-0.2 to 0.2
Primary individual	11	9.1 to 14.4	3.3	0.0*	-0.2 to 0.2
Child	2	1.9 to 2.8	43.5	0.0*	-0.3 to 0.2
Other relative	20	16.5 to 24.1	2.8	0.0*	-0.2 to 0.2
Nonrelative	29	22.4 to 37.7	1.0	0.1*	-0.1 to 0.2
FEMALE					
(Primary) family head	13	11.3 to 15.8	5.2	-0.1*	-0.4 to 0.1
Primary individual	4	3.4 to 5.5	7.5	0.0*	-0.1 to 0.1
Wife	2	1.8 to 2.6	44.6	-0.2*	-0.4 to 0.0
Child	2	1.8 to 2.1	37.8	-0.2*	-0.4 to 0.0
Other relative	18	15.4 to 21.2	3.9	0.3*	0.0 to 0.5
Nonrelative	27	21.3 to 34.9	1.0	0.3*	0.1 to 0.4
EMPLOYMENT STATUS—Con.					
FEMALE, 14 YEARS AND OVER					
Employed in agriculture	63	48.1 to 82.8	0.6	-0.2*	-0.4 to 0.0
Employed in nonagricultural industries	17	15.6 to 18.3	39.2	-1.1	-1.7 to -0.5
Unemployed	65	56.7 to 74.4	2.1	0.0*	-0.3 to 0.4
Not in labor force	19	17.8 to 20.5	58.2	1.2	0.6 to 1.9
WORK EXPERIENCE IN 1969					
MALE, 16 YEARS AND OVER					
50 to 52 weeks	32	30.2 to 33.9	64.3	-6.0	-7.0 to -5.0
48 to 49 weeks	82	74.6 to 90.8	2.6	2.6	1.9 to 3.2
40 to 47 weeks	72	66.4 to 79.2	4.5	2.2	1.5 to 2.8
27 to 39 weeks	69	62.1 to 75.7	4.3	0.7	0.1 to 1.3
14 to 26 weeks	61	55.2 to 68.2	4.8	-0.6	-1.2 to -0.1
13 weeks or less	53	47.4 to 58.6	5.3	0.2*	-0.4 to 0.8
Did not work in 1969	22	20.0 to 24.6	14.2	1.0	0.5 to 1.6
FEMALE, 16 YEARS AND OVER					
50 to 52 weeks	31	28.7 to 32.6	27.7	-5.5	-6.3 to -4.7
48 to 49 weeks	79	69.7 to 88.5	1.6	1.4	1.0 to 1.9
40 to 47 weeks	74	67.3 to 80.7	3.5	1.7	1.2 to 2.3
27 to 39 weeks	65	59.7 to 71.2	5.2	0.2*	-0.4 to 0.8
14 to 26 weeks	57	52.1 to 62.2	6.2	-0.2*	-0.8 to 0.4
13 weeks or less	50	46.2 to 54.5	8.1	0.1*	-0.5 to 0.7
Did not work in 1969	19	17.3 to 20.1	47.7	2.2	1.5 to 2.9
CLASS OF WORKER					
Agriculture:					
Wage and salary	33	26.0 to 42.2	1.2	0.4	0.2 to 0.6
Self-employed	17	12.9 to 21.9	2.3	-0.1*	-0.3 to 0.1
Unpaid family worker	53	27.5 to 100.0	0.1	-0.1*	-0.2 to 0.0
Nonagricultural industries:					
Private wage and salary	16	14.3 to 17.2	72.9	0.5*	-0.1 to 1.1
Government	12	10.9 to 13.9	16.8	-0.2*	-0.6 to 0.2
Self-employed	23	20.3 to 26.8	6.2	-0.3*	-0.7 to 0.1
Unpaid family worker	47	32.7 to 67.7	0.5	-0.1*	-0.3 to 0.0
OCCUPATION					
Professional, technical, and kindred workers	20	18.1 to 22.1	15.0	0.7	0.2 to 1.2
Managers and administrators, except farm	39	35.5 to 42.1	11.6	-2.8	-3.4 to -2.2
Sales workers	30	26.9 to 34.1	5.9	1.1	0.7 to 1.6

MARITAL STATUS					
MALE, 14 YEARS AND OVER					
Married, except separated	2	1.9 to 3.0	70.4	0.0*	-0.3 to 0.2
Separated	43	33.8 to 56.0	1.0	0.1*	-0.2 to 0.3
Widowed	14	10.9 to 19.2	2.5	-0.1*	-0.3 to 0.1
Divorced	24	18.8 to 31.1	1.7	0.4	0.1 to 0.6
Single	2	1.6 to 2.1	24.5	-0.3	-0.5 to -0.1
FEMALE, 14 YEARS AND OVER					
Married, except separated	2	1.7 to 2.6	62.1	0.0*	-0.2 to 0.3
Separated	29	23.7 to 36.3	1.8	0.1*	-0.2 to 0.3
Widowed	6	5.0 to 7.4	11.8	-0.2*	-0.4 to 0.1
Divorced	19	16.0 to 23.3	3.5	0.2*	0.0 to 0.5
Single	2	1.9 to 3.1	20.8	-0.1*	-0.3 to 0.1
EDUCATIONAL ATTAINMENT					
Elementary: 0 to 4 years	34	30.8 to 37.8	3.6	0.2*	0.0 to 0.5
5 years	67	59.3 to 75.3	1.4	0.1*	-0.1 to 0.3
6 and 7 years	46	43.1 to 49.0	6.9	0.6	0.2 to 1.0
8 years	38	36.1 to 40.3	13.3	-0.4*	-0.8 to 0.1
High school: 1 year	47	43.8 to 49.8	7.1	0.4*	0.0 to 0.8
2 years	46	42.9 to 48.4	8.4	0.6	0.1 to 1.0
3 years	48	44.8 to 51.0	6.4	1.1	0.7 to 1.5
4 years	26	25.0 to 27.3	32.8	-2.2	-2.8 to -1.7
College: 1 year	50	45.8 to 53.7	4.2	0.3*	-0.1 to 0.6
2 years	41	37.7 to 44.7	4.8	-0.4	-0.7 to -0.1
3 years	50	44.1 to 55.9	1.7	0.3	0.1 to 0.5
4 years	27	24.2 to 29.4	5.8	-0.8	-1.0 to -0.5
5 years or more	21	18.7 to 24.3	3.5	0.3	0.1 to 0.5
VETERAN STATUS					
Vietnam conflict	22	18.9 to 26.8	5.7	0.8	0.4 to 1.3
Korean conflict	20	16.7 to 23.0	8.2	1.0	0.5 to 1.5
World War II	11	9.6 to 12.9	20.2	0.2*	-0.3 to 0.8
World War I	18	13.4 to 25.0	2.2	0.3*	0.0 to 0.5
Other service	32	27.5 to 36.5	7.1	-0.8	-1.3 to -0.2
Nonveteran	10	8.3 to 10.8	56.6	-1.6	-2.2 to -0.9
EMPLOYMENT STATUS					
MALE, 14 YEARS AND OVER					
Employed in agriculture	26	22.1 to 30.4	4.7	-0.5	-0.9 to -0.2
Employed in nonagricultural industries	15	13.6 to 16.4	68.7	-0.9	-1.5 to -0.3
Unemployed	58	51.1 to 66.1	3.1	-0.2*	-0.6 to 0.3
Not in labor force	17	15.4 to 18.6	23.5	1.5	1.0 to 2.1
Clerical and kindred workers	23	20.7 to 24.6	18.1	0.2*	-0.4 to 0.8
Craftsmen and kindred workers	30	27.4 to 32.6	13.4	0.4*	-0.2 to 1.0
Operatives, except transport.	25	22.3 to 27.0	12.7	0.8	0.3 to 1.3
Transport equipment operatives	26	21.8 to 30.3	3.6	0.4	0.1 to 0.7
Laborers, except farm	49	44.0 to 54.9	4.7	-0.5*	-0.9 to 0.0
Farmers and farm managers	13	9.7 to 17.8	2.3	-0.2	-0.4 to -0.1
Farm laborers and foremen	27	20.2 to 36.7	1.0	0.1*	-0.1 to 0.3
Service workers, except private households	19	17.0 to 21.6	10.1	0.0*	-0.4 to 0.5
Private household workers	15	10.7 to 20.6	1.7	-0.2*	-0.4 to 0.0
INDUSTRY					
Agriculture, forestry, and fisheries	14	11.1 to 17.4	3.7	0.1*	-0.1 to 0.4
Mining	27	18.9 to 38.2	0.7	0.1*	0.0 to 0.3
Construction	21	18.3 to 24.6	5.9	-0.1*	-0.4 to 0.3
Manufacturing	14	13.2 to 15.9	27.8	-1.4	-1.9 to -0.8
Transportation, communications, and other public utilities	14	11.4 to 16.0	7.1	-0.1*	-0.4 to 0.3
Wholesale and retail trade	19	17.4 to 21.0	18.2	1.2	0.7 to 1.8
Finance, insurance, and real estate	11	8.8 to 13.5	5.4	-0.1*	-0.3 to 0.2
Business and repair services	34	29.3 to 40.6	2.7	0.5	0.1 to 0.8
Personal services	15	11.9 to 17.8	4.7	-0.3	-0.6 to -0.1
Entertainment and recreation services	38	27.2 to 51.8	0.6	0.1*	-0.1 to 0.2
Professional and related services	10	8.8 to 11.5	17.3	-0.1*	-0.5 to 0.2
Public administration	15	12.3 to 17.5	5.9	0.0*	-0.3 to 0.3
PERSONS INCOME IN 1969					
Without income	25	24.1 to 26.9	22.1	1.4	0.9 to 1.9
Loss	78	55.6 to 100.0	0.2	-0.1	-0.2 to -0.1
$1 to $999	47	45.0 to 49.3	14.2	-1.1	-1.6 to -0.6
$1,000 to $1,999	51	48.0 to 53.5	9.7	-0.2*	-0.7 to 0.3
$2,000 to $2,999	55	51.6 to 58.2	7.0	-0.2*	-0.6 to 0.2
$3,000 to $3,999	55	51.5 to 58.4	6.3	0.1*	-0.3 to 0.5
$4,000 to $4,999	53	49.2 to 56.4	5.7	-0.1*	-0.5 to 0.3
$5,000 to $5,999	55	51.1 to 58.5	5.4	0.2*	-0.2 to 0.6
$6,000 to $6,999	55	51.4 to 59.1	5.1	-0.1*	-0.4 to 0.3
$7,000 to $7,999	54	50.0 to 57.9	4.8	-0.2*	-0.6 to 0.1
$8,000 to $8,999	53	49.2 to 57.7	4.0	0.0*	-0.3 to 0.4
$9,000 to $9,999	53	48.1 to 57.6	3.1	0.1*	-0.2 to 0.4
$10,000 to $14,999	35	32.7 to 37.6	8.3	-0.2*	-0.6 to 0.2
$15,000 to $24,999	37	33.1 to 41.0	3.1	0.2*	-0.1 to 0.4
$25,000 or more	40	33.8 to 47.8	0.9	0.3	0.1 to 0.4

* Indicates that the net difference rate is not significantly different from zero at the 95 percent confidence level.

1/ A positive value means there were more persons in the census than in the CPS in a given category. A negative value means there were fewer persons in the census than in the CPS in a given category.

provides an approximate measure of the distorting effect each variable has upon the cross-classification. If, then, any of the characteristics have a high index value, the cross-classification may be seriously distorted. The net difference rate estimates the level of bias in the particular census distribution, where the CPS classification is assumed to be more accurate, and is simply the difference between the census and CPS estimates of the proportion of persons in a given category. For this study, the assumption of greater accuracy in CPS is not necessarily true for some characteristics. Therefore, the net difference rates do not always estimate error but merely differences in results obtained from the two data collection systems.

In general, the basic demographic and social characteristics (age, sex, race, household relationship, and marital status) exhibit a high level of response or classification consistency between the CPS and Census, as can be seen in table 1. For each, the estimated L-fold index of inconsistency was under 20 and, with the exception of the age classification for Negroes, the L-fold indices were under 10. These relatively low indices were observed for both men and women for each characteristic and for age classification by major residence categories (i.e. metropolitan, central city; metropolitan, outside central city; and nonmetropolitan, urban and rural). Also, none of the estimated indices differed appreciably from those ascertained in the 1960 CPS-Census matching study.

In table 2, the estimated indices of inconsistency and net difference rates are shown for each category in the distributions. Here again the age classification exhibits a high level of consistency as none of the indices for five-year age groups exceed 20 and most are under 10. Furthermore, the net difference rates indicate there are no substantial biases in the age classification.

Classification of the population by race is also highly consistent for whites and Negroes but slightly more inconsistent for persons of other races. Also, the net difference rates indicate a slight downward bias in the census classification of persons as white and a small upward bias toward classifying perions in races other than white or Negro.

Although the L-fold indices of inconsistency for household relationship and marital status are quite low, some specific categories within these distributions are somewhat more inconsistent. The relationship categories "other relative" and "nonrelative" have indices which approach or exceed 20. The classification "other relative" is one which would, most likely, be more accurately determined by an experienced CPS interviewer. However, the classification as a "nonrelative" could easily change from one enumeration to another. In households not occupied by members of the same family the person being interviewed or completing the questionnaire would be classified as a primary individual and all other persons in the household would be "nonrelatives." Obviously, the person answering the questions could be different between the CPS and census and the classification would, therefore, vary. With regard to the marital status classification, the categories "separated" and "divorced" were moderately inconsistent between the CPS and census. Finally, the net difference rates for both marital status and household relationship indicate that all categories are relatively free of bias.

Classification of persons by educational attainment was moderately inconsistent between the CPS and census. The L-fold indices are in the high thirties for both men and women. However, classification at the terminal levels of education, that is 4 years of high school, 4 years of college, and 5 or more years of college, was more consistent than was true for other attainment levels. Specifically, the indices for these categories were between 20 and 30. As is evidenced by the net difference rates, differences between the CPS and census classification by educational attainment are largely offsetting and, hence, little bias is noted. There is some indication, however, that the census may tend to understate slightly the terminal education categories relative to CPS. However, it may not be appropriate to view these differences as an indication of error in the census education statistics. There is some speculation that a respondent in a personal interview situation (as in CPS) may be more likely to erroneously report education at a terminal category than is true when the person is actually completing a questionnaire.

The estimated L-fold index of inconsistency for veteran status and period of service (15) reveals a low level of inconsistency between the two data sources. The reporting of veteran-nonveteran was relatively uniform as was the classification of World War II veterans. The remaining service categories, however, have index values near or above 20, and the estimated index for the residual "other service" category exceeds 30. As seen from the net difference rates, the census tends to understate the "other service" and "nonveteran" categories relative to CPS and to slightly overstate the proportion of veterans who served in the Vietnam and Korean conflicts.

The economic characteristics (i.e. employment status, work experience, class of worker, occupation, industry, and income) were, in general, less consistent in classification between the CPS and Census than was true for the demographic and social variables. This same relationship was observed in the 1960 matching study. In particular, the occupational classification for men and the distribution by weeks worked and most types of income for both men and women had estimated indices ranging from the low thirties to over fifty. On the other hand, the L-fold indices for employment status, class of worker, and industry with values of 20 or less indicate a relatively low level of inconsistency.

The indices of inconsistency for employment status cannot be strictly interpreted as measures of response or classification error. Since many persons could have experienced a change in employment status between the March 1970 CPS interview and the time of census enumeration, some portion of the difference in classification is valid. The indices reflect a combination of classification

errors and actual changes, and the index of inconsistency is, therefore, overstated to the extent that actual changes occurred. Even so, the L-fold index of about 20 for men and women indicates a reasonably high level of consistency. Among the four employment status categories, classification as "employed in nonagricultural industries" and as "not in the labor force" was fairly uniform between CPS and Census. However, the indices associated with the category "employed in agriculture" were somewhat higher, especially for women; and the classification "unemployed" was very inconsistent for both men and women. It must be remembered, though, that unemployment is subject to change over a short period of time and most of the response differences observed may reflect real changes.

The L-fold index of inconsistency on work experience for both men and women (43 and 37 respectively) reflects a moderately high level of disagreement between the CPS and census classification. However, the indices for specific categories indicate that the dichotomy of worked in 1969/did not work in 1969 and the identification of year-round workers (50 to 52 weeks) were fairly consistent. On the other hand, the identification of specific weeks worked categories for other than year-round workers was highly variable. The net result of differences in classification indicates that the census tended to understate the proportion working 50 to 52 weeks and to overstate the proportion who worked from 40 to 49 weeks and the proportion who did not work in 1969.

Of the three census "job content" classifications, class of worker and major industry group were reasonably consistent with the CPS. Among the seven class of worker categories, three had estimated indices under 20. However, the classification of unpaid family workers and of agricultural wage and salary workers was less uniform. All but four of the twelve major industry categories had estimated indices of inconsistency under 20 and none of the estimated indices exceeded 50. However, the net difference rates indicate a slight understatement of the proportion working in manufacturing industries in the census and a small overstatement for wholesale and retail trade.

The L-fold indices for major occupation (31 for men and 21 for women) reflect a moderate level of inconsistency in classification. For only three groups--farmers and farm managers, service workers, and private household workers--were the estimated indices below 20. In seven other groups the indices ranged from the low twenties to the low thirties. However, for two occupations--managers and administrators and laborers--the indices were about 40 and 50, respectively. Moreover, for managers the net difference rate reflects an understatement of about 3 percentage points in the census. However, evidence has indicated that the CPS occupation item, prior to a revision instituted in December 1971, tended to overestimate the proportion of managers. Since the comparisons made in this study are based on the old CPS occupation item, it may not be appropriate to view the census count of managers as negatively biased. For a number of occupation descriptions, the determination of major occupation group is very difficult. For instance, the distinction between a warehouseman (a laborer) and a fork-life operator (an operative) is often very subtle, based on the information at hand. As a result, there historically has been a great deal of inconsistency in classification especially among the "blue-collar" occupation groups. Since the laborer group constitutes a much smaller proportion of employed persons than do craftsmen and operatives, differences in classification among these groups have relatively more effect on the index for the smaller group.

Gross differences in income classification between the CPS and census resulted in a rather high level of inconsistency. The estimated value for the L-fold index of inconsistency for persons total income was 45 (50 for men and 43 for women). Among the type of earnings categories, wage or salary income was classified with a moderate level of inconsistency, as reflected by indices of inconsistency in the thirties. However, the classification by nonfarm and farm self-employment income was highly inconsistent. The type of "income other than earnings" which was most consistently reported was Social Security income. The limited range in the amount of income that can be collected under the Social Security system may explain the relatively low levels of inconsistency. The other sources of unearned income, including public assistance and "all other income," were characterized by high levels of inconsistency. The census category "all other income" was broken down into three separate questions in the CPS and the extra questionnaire detail may have helped the respondent in the CPS to recall small amounts of income from relatively unimportant sources.

The gross errors or differences in classification between the CPS and census seem to be largely offsetting since, at least for total income, there is little or no bias associated with the specific income categories. However, the proportion of persons reporting "no income" was overstated slightly in the census and the proportion reporting income under $1,000 was understated somewhat. This may again signify that respondents were more likely to recall small amounts of income in the CPS than they were in the census.

For the second portion of the study, a matching of Census and IRS data was accomplished by attempting to obtain 1969 Federal income tax returns for all individuals 14 years old and over in the CPS-Census sample. The 1970 Census is the third Census to be evaluated by using tax return data as an income benchmark[1]. The strict confidentiality of both the Census and IRS data make this type of matching evaluation difficult. All matching work was done by the Bureau of the Census in order to preserve the confidentiality of replies to census questions; no one other than Bureau employees, who are sworn to uphold the confidentiality of all Census information, had access to the information. The confidentiality of IRS records was also safeguarded. A detailed description of the procedures implemented to safeguard the confidentiality of the Census, CPS, and IRS data will be published by the Bureau in the final report on this project.

The Social Security Number (SSN) for persons 14 years old and over has been collected as part of the March Current Population Survey (CPS) for several years. These SSN's as well as name and address were used by Census Bureau employees to obtain tax returns. Once a tax return was located, the CPS identifying information was assigned to the IRS data and the name and SSN were not used in the actual match of the data.

For the cases the Census agents could not find, the majority were because no SSN was available from the CPS. For these additional efforts were made to obtain respondents' social security numbers from the Social Security Administration by name and birth date search of Social Security records. This procedure resulted in obtaining an additional 500 tax returns from the 2,000 needed bringing the final total to 8,434 returns located.

The assumption of the study is that Internal Revenue Service (IRS) income data is more accurate than Census income data. Since the definition of taxable income is different from the definition of total money income as used in the census, the comparison is not as straightforward as is often thought. In addition, not all persons are required to file tax returns, especially low-income persons and persons living on transfer payments. Consequently, the study is limited to persons filing returns located in the CPS or Census samples.

Six questions on income were asked in the 1970 Census. Listed below are the definitions of the income types from both sources. Wages and salaries is the only income type with the same definition in both Census and IRS. Other problems: (1) There was some tendency to report second job wages, in miscellaneous income (Schedule E) rather than wages and salaries, but this was limited to small amounts, (2) the use of Schedule F and C net income to approximate self-employed farm and nonfarm self-employment income is very rough because of the impossibility of splitting partnership income into farm and nonfarm sources, (3) the reporting of some sale of livestock as capital gains (Schedule D), and (4) the IRS rules defining current year expenses for farm income.

Census Income Type	IRS Income Type
Total Money Income (TMY)	Adjusted Gross Income (AGI)
1. Wages, salary, commissions, bonuses, and tips	Wages and Salaries
2. Earnings from nonfarm business, professional practice or partnership	a. Net income from Schedule C b. Partnership income from Schedule E 2/
3. Earnings from own farm	a. Net income from Schedule F 2/ 3/
4. Social Security or Railroad Retirement income	Not reported to IRS
5. Public Assistance or Welfare payments	Not reported to IRS
6. Other source - Interest, dividends, veterans payments pensions, and other regular payments	a. Schedule B - Gross Dividends and Interest 4/ b. Schedule E - Pension annuity net rent, net royalties, income from estates or trust, small business dividends, misc. income

The census data was tabulated for persons, family, and unrelated individuals. Tabulations from the matched Census and IRS file by persons is not possible because income on joint returns cannot be assigned separately to either the husband or the wife. The possibility of splitting wages and salaries reported on joint returns using W-2 Forms was considered, but the W-2 Forms were not available for many of the returns.

Although the data have been tabulated several ways, the data used in this paper are restricted to a matched household head, and household head and wife matched to either a joint return or separate returns. The data are preliminary.

As table 3 shows, the census total money income (TMY) was only 2.9 percent less than IRS's adjusted gross income (AGI). If the conceptual differences are taken into account as far as possible by removing capital gains from AGI and removing transfer payments (Social Security and public assistance) from Census total money income the Census aggregate is 3.4 percent less than the IRS aggregate. The consistency as measured by the L-fold index is high, 66.3 and 63.6 (see table 3).

Wages and salaries is the only type of income with an L-fold index below 50 when computed on the attached income intervals. Even when the classes are broadened to $5000 intervals, the inconsistency remains high at 33 (see table A). The Census captured 95.0 percent of the wages and salaries reported to IRS.

The L-fold indices for nonfarm and farm self-employment income were 69.8 and 77.0 percent respectively. The aggregate census farm income was over twice the net IRS farm income reported on Schedule F, but this is partly due to conceptual differences.

Table A.—NUMBER OF MATCHED HEADS OR MATCHED HEAD AND WIFE BY WAGES AND SALARIES INCOME IN THE 1970 CENSUS AND 1969 TAX RETURNS

IRS Wages or Salaries	Total	Under $5000	$5000 to $9999	$10,000 to $14,999	$15,000 and over
Total	6175	2128	2028	1309	709
Under $5,000	2043	1662	254	81	46
$5,000 to $9,999	1984	280	1513	157	34
$10,000 - $14,999	1395	131	200	980	84
$15,000 and over	752	55	61	91	545

L-fold index is equal to 33.2.

Summary

The small differences in Census and IRS aggregates for total money income, and wages and salaries are encouraging. However, the high inconsistency, even from wages and salaries, indicates there is still a lot of work to be done in improving the collection of all sources of income data.

FOOTNOTES

1/ The accuracy of income data collected in both the 1950 and 1960 Census was evaluated by using tax return data. The results of the 1960 effort were published in Record Check of Accuracy of Income Reporting, Series ER-60, No. 8, U.S. Bureau of the Census. The results of the 1950 study were published in An Appraisal of the 1950 Census Income Data, "Income Reported in the 1950 Census and on Income Tax Returns," Herman P. Miller and Leon R. Paley.

2/ Partnership income on Schedule E could be from farming but there is no way to separate partnership income by source.

3/ Some capital gains income (Schedule D) resulting from sale of livestock would also be considered farm income, but this cannot be easily identified.

4/ State tax refunds are supposed to be reported as interest, but interest identified as tax refunds was excluded from interest.

Table 3.—NUMBER OF MATCHED HEADS OR MATCHED HEAD AND WIFE BY TYPE OF INCOME FOR 1969 (PRELIMINARY)

Income Class Intervals	Total Income		Adjusted Total Income		Wages and Salaries		Nonfarm Self-Employment		Farm Self-Employment	
	Total Money Income (TMY)	Adjusted Gross Income (AGI)	TMY (Minus) (SS+PA) Transfers	AGI (Minus) Capital Gains	Wages and Salaries	Wages and Salaries	Nonfarm Self-Employment Income	IRS Net from Schedule C and Partnership income from Schedule E	Farm Self-Employment Income	IRS Net from Schedule F
	(Census)	(IRS)	(Census)	(IRS)	(Census)	(IRS)	(Census)	(IRS)	(Census)	(IRS)
Total Matched Units.....	6,174	6,174	6,174	6,174	6,174	6,174	6,174	6,174	6,174	6,174
LOSS	10	--	11	16	--	--	34	153	32	133
None.....	47	56	208	48	1,021	811	5,542	5,405	5,890	5,813
$1 - $999	124	170	191	180	172	254	86	140	65	80
$1,000 - $1,999	263	276	264	291	211	228	61	70	44	41
$2,000 - $2,999	286	318	275	325	200	208	41	52	31	28
$3,000 - $3,999	375	322	296	319	238	245	52	46	14	14
$4,000 - $4,999	359	372	341	372	286	297	33	28	13	12
$5,000 - $5,999	407	390	399	389	334	330	53	42	22	7
$6,000 - $6,999	481	392	441	390	436	379	38	30	12	14
$7,000 - $7,999	487	455	466	447	445	417	24	30	6	11
$8,000 - $8,999	500	491	489	499	431	431	22	26	15	6
$9,000 - $9,999	423	461	417	447	382	427	21	13	9	2
$10,000 - $11,999	742	733	725	734	702	682	34	23	12	7
$12,000 - $14,999	707	755	694	758	607	713	37	25	6	4
$15,000 - $24,999	747	781	744	770	597	649	40	44	2	1
$25,000 AND OVER	216	202	213	189	112	103	56	47	1	1
Median (dollars)1/.....	8,496	8,684	8,398	8,621	7,424	7,803	--	--	--	--
Aggregate income (dollars)...	60,066,846	61,882,002	58,578,912	60,653,376	48,027,546	50,552,712	5,291,118	4,568,760	913,752	407,484
Net percent difference2/.....	-2.9	(X)	-3.4	(X)	-5.0	(X)	+15.8	(X)	+124.2	(X)
L-fold Index of inconsistency.....	66.3		63.6		47.3		69.8		77.0	
95% confidence interval on index of inconsistency.....	65.0-67.7		62.3-64.9		45.7-48.9		65.0-75.8		70.4-86.1	

NA Not applicable. -- Equals zero.

1/ The none's were included in the computation of the median.

2/ $\frac{\text{Census-IRS}}{\text{IRS}} \times 100$ = Net percent difference. The 95 percent confidence intervals for the net percent difference are: Total income -5.5 to -0.3, adjusted total income -6.6 to -0.2, wages and salaries -6.0 to -4.0, nonfarm self-employment -4.5 to 36.1, and farm self-employment 37.1 to 211.3.

ENUMERATOR VARIANCE IN THE 1970 CENSUS

Benjamin J. Tepping
and
Barbara A. Bailar, Bureau of the Census

1. Introduction

Many papers over the last 30 years have been written on the subject of interviewers as a source of error in survey data. Yet few large-scale experiments have been conducted which give precise measures of this interviewer effect. Some efforts which have been made were reported by Fellegi (1) on an experiment carried out by Statistics Canada, by Kish (4) on an experiment carried out at the University of Michigan, and by the U.S. Bureau of the Census (6).

The Bureau of the Census has had experiments to measure the enumerator (interviewer) effect in the last three censuses of population and housing. Hanson and Marks (3) reported on the 1950 Enumerator Variance Study which took place in 21 purposively selected counties in Ohio and Michigan. The results of that study showed that the variability in census statistics which could be attributed to interviewers was, on the average, roughly equal to the sampling variability one might expect from a 25-percent sample of the population.

The results of the 1950 experiment greatly influenced the Census Bureau to introduce the use of self-enumeration on a widespread basis in the 1960 Census. To find out whether the increased use of self-enumeration had any effect on the level of enumerator variability, a large-scale experiment was conducted as part of the 1960 evaluation and research program. The results of that study showed that the level of variability in census statistics accounted for by enumerators in 1960 was reduced to about one-fourth of the 1950 level. However, even in 1960 enumerator variability had a considerable impact on statistics for small areas.

In 1970, some important changes were made in the census-taking procedures. The United States was divided into three kinds of areas, in each of which a different kind of census-taking procedure was used. The large central city areas were enumerated by a "centralized mail" procedure. The less densely settled areas of the country were enumerated by a "conventional procedure". The remainder of the country, containing about 50 percent of the population, was enumerated by a "decentralized mail" procedure.

The 1970 Enumerator Variance Study was confined to the decentralized mail areas on the ground that this type of census-taking procedure was most likely to be followed in the future. Thus, the estimates of enumerator variability presented in this paper are applicable to decentralized mail areas only.

Within the decentralized areas, the field procedure was as follows: An Address Register, a list of housing unit addresses to which blank questionnaires had been mailed, was supplied to the enumerator for an enumeration district (ED). The enumerator was to see that a completed census questionnaire was returned for each address listed in the Address Register for an ED. Enumerators were instructed in four areas: (1) how to check-in forms received by mail; (2) how to edit the short-forms (100 percent census schedules); (3) how to edit the long-forms (the sample census schedules); and (4) how to followup for nonresponses and inconsistencies.

As we shall show later, there was considerable variability among the enumerators in the interpretation of the editing rules. The possibility that this change in the enumerator's editing role might affect the level of enumerator variability was a primary consideration in deciding to carry out another enumerator variance study as part of the 1970 Census.

2. The Design of the Enumerator Variance Study

Of the 167 census district offices established in which a decentralized mail procedure was used a probability selection of 35 was made in which to carry out the Enumerator Variance Study (EVS). In each area two crew leader districts were selected, within which the enumerator assignments were grouped into clusters of four. All clusters within the two selected crew leader districts were included in the study. Altogether 259 of these clusters of enumerator assignments were included in the study.

The assignments of four enumerators were interpenetrated in each of these clusters. The listings within an Address Register were randomized such that within every group of eight listings designated to receive a long-form two were assigned to each of the four enumerators.

As far as possible, the procedures followed in the EVS areas were exactly the same as those followed in other decentralized mail areas. The major exceptions were:

1. A statistician was assigned to each EVS district office to supervise the study.

2. The enumerators were paid higher piece rates in EVS areas since they would travel more.

3. The enumerators and crew leaders received additional training on EVS procedures.

4. The enumerators worked with copies of the Address Registers while the crew leaders kept the original.

3. Processing of the Data

As with the regular census materials, all EVS census materials were sent to the Jeffersonville,

Indiana office for processing. The census schedules were coded, microfilmed, converted to magnetic tape, edited, and the final census data tabulated. In all of this processing, the EVS schedules were treated just as any other schedules. Thus, the results of this study are applicable to the final published census statistics.

The data for the ED's included in the EVS were copied from the census sample tapes for the States in which the EVS offices were located. A basic identification record was made for each address in each ED showing the State, ED number, cluster number, group-of-eight number, and interviewer number. These identification records were matched against the census tapes so that enumerator assignments could be identified in the computations. The matching operation left only 1.6 percent of the originally randomized addresses unmatched. This was an improvement over the 1960 matching operation when we were unable to match 13 percent of the originally randomized units.

At the completion of the matching operation, the EVS sample contained approximately 127,000 housing units and 378,000 persons. This was slightly larger than the sample for the 1960 study which contained about 122,500 housing units and 370,000 persons.

4. The Estimation Procedure

The mathematical model used in this study is the model described by Hansen, Hurwitz, and Bershad (2). The basic assumption in this model is that a response from a given unit of the population is a random variable from a probability distribution. Thus, if it were possible to record responses for each individual repeatedly, a distribution of responses for each unit would be generated. In a census or survey, we first sample respondents and then sample from the distribution of possible responses for each sample person.

The survey process is regarded as being repeatable. Each survey is a trial from all possible repetitions of the survey process under the same general conditions which include the auspices of the survey, the questions used, the method of recording and processing responses, and the general social environment.

If one is interested in estimating a mean, $\bar{x}$, for an area the size of one enumerator's assignment, then it can be shown that under certain assumptions the mean-square error of that mean may be written as:

$$MSE(\bar{x}) = \frac{1}{n}\left\{\sigma_R^2[1+(n-1)\rho_R] + \frac{N-n}{N-1}\sigma_s^2+2(n-1)\sigma_{Rs}\right\} + B^2 . \quad (1)$$

In this expression N is the number of sampling units in the population, and n is the number of sampling units in the sample. The simple response variance, σ_R^2, is the trial-to-trial variability of response for a given unit. The sampling variance, σ_s^2, is the variance of the mean response of the units in the population. The square of the bias is denoted by B^2. The term $\rho_R\sigma_R^2$ is the correlated component of response variance, measuring the contribution to total variability caused by the correlation, ρ_R, of response deviations within trials. One reason for a correlation among response deviations is that one enumerator may interpret a question differently from other enumerators. His tendency to accept nonresponses, his possible misunderstanding of the training on certain questions, and other such factors tend to cause a positive correlation in the response deviations within his assignment. The correlations can also be introduced by crew leaders and by coders. The EVS is designed so that this term measures the correlations induced by the enumerators only. The remaining term in equation (1) is σ_{Rs}, the covariance between response and sampling deviations, often assumed zero.

One reason for using interpenetrating subsamples is to estimate the correlated component of response variance. Of each group of eight successive listings in each of the 259 clusters, two listings were assigned to each of the four enumerators.

There were N listings, of which 20 percent were included in the census sample, the sample total designated by n. The sample listings were assigned to enumerators in groups-of-eight. We index groups-of-eight by the subscript g, and denote by b the number of groups-of-eight in the area. Within each group-of-eight there were $\bar{N}$ listings.

Within a group-of-eight sample listings, each enumerator was assigned $\bar{n} = 2$ listings. The subscript j indexes the listings within a group. The subscript h indexes the enumerators working in the cluster and the range is from 1 to k where k = 4. Thus, the recorded response for a given characteristic for the j-th unit in the g-th group-of-eight assigned to the h-th enumerator in the t-th trial of the process is denoted by x_{htgj}. From the data available from the EVS, there are five sample sums of squares and cross-products which can be used in estimation. These are:

$$S_0 = \frac{1}{kb\bar{n}}\sum_{h=1}^{k}\sum_{g=1}^{b}\sum_{j=1}^{\bar{n}} x_{htgj}^2$$

$$S_1 = \frac{1}{kb\bar{n}(\bar{n}-1)}\sum_{h=1}^{k}\sum_{g=1}^{b}\sum_{j\neq j'}^{\bar{n}} x_{htgj}x_{htgj'}$$

$$S_2 = \frac{1}{kb(b-1)\bar{n}^2}\sum_{h=1}^{k}\sum_{g\neq g'}^{b}\sum_{j,j'}^{\bar{n}} x_{htgj}x_{htg'j'}$$

$$S_3 = \frac{1}{k(k-1)b\bar{n}^2}\sum_{h\neq h'}^{k}\sum_{g=1}^{b}\sum_{j,j'}^{\bar{n}} x_{htgj}x_{h'tgj'}$$

$$S_4 = \frac{1}{k(k-1)b(b-1)\bar{n}^2}\sum_{h\neq h'}^{k}\sum_{g\neq g'}^{b}\sum_{j,j'}^{\bar{n}} x_{htgj}x_{h'tg'j'} .$$

An estimator of the total variance which we used in this study is:

$$\hat{Var}(\bar{x}_t) = \frac{1}{b\bar{n}}S_0 + \frac{\bar{n}-1}{b\bar{n}}S_1 + \frac{b-1}{b}S_2 - \frac{1}{b}S_3 - \frac{b-1}{b}S_4 . \quad (2)$$

An estimator of the sampling variance is:

$$s_{\bar{x}_t}^2 = \frac{1}{b\bar{n}}S_0 - \frac{1}{b\bar{n}}S_1 .$$

The difference of these two estimators, an

estimator of the correlated component of response variance, is:

$$u^2 = \frac{1}{b}(S_1 - S_3) + \frac{(b-1)}{b}(S_2 - S_4). \qquad (3)$$

Many of the statistics are in the form of ratios. For example, one may be interested in the proportion of persons 16 years of age and over who are employed as service workers. Let y denote service workers and x denote persons 16 years of age and over. Then, we estimate

$$\frac{y}{x} = \frac{\sum_{h}^{4}\sum_{s}^{b}\sum_{j}^{a} y_{htsj}}{\sum_{h}^{4}\sum_{s}^{b}\sum_{j}^{a} x_{htsj}} . \qquad (4)$$

An approximation to the relative variance of a ratio, y/x, is:

$$V^2_{y/x} = V^2_y + V^2_x - 2V_{xy} \qquad (5)$$

where V_y^2 is the relative variance of y, V_x^2 is the relative variance of x, and V_{xy} is the relative covariance of x and y. A consistent estimate of the relvariance of a ratio is:

$$v^2 = \frac{u_y^2}{(Ey)^2} + \frac{u_x^2}{(Ex)^2} - \frac{2u_{xy}}{(Ey)(Ex)} \qquad (6)$$

where u^2 is of the form shown in equation (3).

Estimates of the numerators and denominators of the terms shown in equation (6) were computed for each of the 259 clusters. Numerators and denominators were each weighted and averaged over the clusters, and the weighted figures were substituted in equation (6).

5. Results

A. Comparison with 1960 Results

The main result of the study can be stated simply: the level of enumerator variability in the 1970 census is at least as high as the level in the 1960 census.

We make the 1960-1970 comparison by comparing response relvariances for identical items in the two censuses. We can compare these relvariances directly since the estimates in each case were for an area enumerated by one enumerator. For this comparison, the correlated components were estimated by use of equation (3) to keep the estimation procedure identical with that used in the 1960 census. Table 1 shows this comparison for the 82 items for which the correlated component of response relvariance was estimated in both 1960 and 1970. For 43 of these items, the 1970 response relvariances were larger in 1970 than in 1960 and for 39 they were smaller.

However, one is not usually interested in looking at statistics for an area of the size that could be enumerated by one enumerator. Rather, one is more interested in statistics for blocks, tracts, or States which are usually based on the work of several enumerators. The size of an enumerator's area in 1970 was about twice as large as the size in 1960. Thus, even if the level of variability in 1970 was the same as in 1960 for an area which was enumerated by one enumerator, the level of variability for tracts and other larger areas would be about twice as large as in 1960.

To make the comparison more meaningful, one should look at the ratio of 1970 to 1960 variances by classes of items. Then, we see that for items concerned with payment of utilities for rented units, the correlated response variance was smaller in 1970 than in 1960. We notice also that though the response relvariances for "not reported" items were large in 1970, they were about half of the size that they were in 1960.

For nativity items, the 1970 relvariances were at least twice as large as the 1960 relvariances. Also for the characteristic "residence 5 years previous to the census", we see an increase in the 1970 response relvariances. This was a complex item for respondents and interviewers in 1960 as well as 1970.

The educational attainment items show a somewhat mixed pattern. Of nine categories, six showed larger response relvariances in 1970 than in 1960. For some of these categories, the increases were substantial. The school enrollment items when defined as actual year in which enrolled show larger relvariances in 1970 than in 1960 except for college years. The kind of school in which enrolled shows larger 1970 relvariances for public school enrollments and smaller 1970 relvariances for private school enrollments.

Response relvariances were larger in 1970 than in 1960 for three of four categories for number of children ever born.

There are only two labor force items which were studied in both 1960 and 1970. One shows a smaller relvariance and one shows a larger relvariance.

We see a somewhat mixed pattern for occupation items. Four of the seven categories have larger 1970 relvariances.

The pattern for the three kinds of income items is also mixed. One consistent note throughout the three types of income was that for the category "males, $5,000 to $6,999". For wage and salary income, the ratio of 1970 to 1960 relvariances was 2.2; for self-employment income, the ratio was 2.9; and for income other than earnings, it was 2.3.

The single veteran status item had a relvariance three times as big in 1970 as in 1960.

These results are preliminary. We have much more work to do in arriving at a statement on the overall level of response variability in the 1970 census as compared with the 1960 census. However, the relvariances shown in Table 1 give the impression that the level of response variability in 1970 is at least as large as the 1960 level, if not larger. How do we account for this? The change in census-taking procedures that would have had the greatest impact on the enumerators was their editing function. In 1960, enumerators copied the entries made by the householders or which they themselves had made on the household questionnaires to another form which could be machine

processed. This transcription operation, though it provided an opportunity for copying error, forced the enumerators to review the questionnaires. In 1970, the original entries were made on forms which could be machine processed. The enumerators were instructed in editing procedure.

As the result of applying the editing instructions, an enumerator could judge a form to be complete, or could judge a form to fail the edit, in which case he would have to contact the unit for additional information. The contact could be by telephone or in person. The instructions told how to make the decision on whether to followup by telephone or personal visit.

During the course of the census, it was important to have accurate information about the percentage of census questionnaires returned by mail, about the failure rates from editing short- and long-forms, and about the size of the followup assignments. Each enumerator filled out a form giving such information. We examined these forms for EVS ED's, computed the mail-return rates, the short- and the long-form edit-failure rates for EVS offices and compared them with the rates for all decentralized mail areas. These rates were as follows:

	All decentralized offices	EVS offices
Mail-return rates	.81	.83
Short-form edit-failure rates	.13	.12
Long-form edit-failure rates	.43	.47

The rates were comparable as was expected. We now wanted to estimate the variability among enumerators in the edit-failure rates. Tepping (5) developed a model to estimate the overall variability in the rates and the variability attributable just to enumerators. It is this latter part of the variability which accounts for the non-uniformity of the application of the editing rules. He found that the average s_r^2, the part of the total variance accounted for by enumerator variability was .00356 for the short-form edit-failure rate and .02370 for the long-form edit-failure rate. Thus, a long-form edit-failure rate of .47 has a standard error of .15. This means that there was a considerable amount of variability among the enumerators in the application of the editing rules.

B. Levels of Variability in 1970 Census Statistics

Tables 2 and 3 show detailed results on response and sampling relvariances for a small number of all the 1970 Census statistics that were studied. We selected a number of characteristics over a broad range which might be of interest before we looked at the results. The selected characteristics included some items contained on the 100 percent census schedules as well as the sample schedules. The complete description of the results for all characteristics studied will be issued in two evaluation reports sometime in 1974. Table 2 shows results for selected housing items; Table 3 shows results for selected population items. Almost all of these items were ratios. Thus, equation (6) was used to compute the estimates.

These estimates apply to an enumeration by one interviewer in an area having about 2,500 housing units and 7,500 persons. To determine the response relvariance for areas having more than 7,500 persons, the response relvariances must be divided by N/7,500 where N is the population in the area of interest.

Table 2 shows that the ratios of response to sampling variability for duration of vacancy for vacant units were over 1.0 for all categories. This is an item for which the enumerator would have had to followup, since no schedules would have been returned for vacant units. Thus, the enumerator variability exhibited by this set of items shows no gain due to the increased use of self-enumeration.

Another interesting result in Table 2 is the ratio for 1-room units of 1.51. This probably relates to a difficulty among enumerators in the classification of efficiency apartments.

The first four items shown in Table 3 for population items were 100 percent items and, in the complete census, would not be subject to sampling variability. We would usually consider them to be exact except for simple response variances. Thus, the response relvariance shows the amount of variability which should be used for these items for an area of one enumerator assignment. The sampling relvariances shown are those that are applicable to the proportions which we estimate using the 20-percent sample.

For most of the characteristics shown in Table 3, we see that the ratios of response to sampling variance are usually below .50. Only the nonresponse categories show ratios over 1.0.

The results presented above give some indication that enumerator variability was still a problem in the 1970 Census. The level of variability increased over 1960 for some items and decreased for others. For some items, the variability attributable to enumerators was an important part of the total variability of census statistics.

REFERENCES

1. Fellegi, I.P. "Response Variance and Its Estimation". Journal of the American Statistical Association, Vol. 59, 1964, pp. 1016-1041.
2. Hansen, M.H., Hurwitz, W.N., and Bershad, M.A., "Measurement Errors in Censuses and Surveys". Bulletin of International Statistical Institute, Vol. 38, Part 2, Tokyo, 1961, pp. 359-374.
3. Hanson, R.H., and Marks, E.S., "Influence of the Interviewer on the Accuracy of Survey Results". Journal of the American Statistical Association, Vol. 53, 1958, pp. 635-655.
4. Kish, Leslie, "Studies of Interviewer Variability for Attitudinal Variables". Journal of the American Statistical Association, Vol. 57, 1962, pp. 92-115.
5. Tepping, B.J., Variability in Edit-Failure Rates, Unpublished Bureau of the Census memorandum, September 15, 1970.
6. U.S. Bureau of the Census, Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Effects of Interviewers and Crew Leaders, Series ER60, No. 7, Washington, D.C., 1968.

TABLE 1.--COMPARISON OF CORRELATED COMPONENT OF RESPONSE RELVARIANCES FOR IDENTICAL ITEMS: 1960 AND 1970 CENSUSES

Characteristic	Response relvariances		Ratio of 1970 to 1960 relvariances
	1960 (1)	1970 (2)	(3)
Rented housing units paying for:			
Electricity	.00379	.00151	0.4
Gas	.00848	.00322	0.4
Water	.12531	.00555	0.0
Fuel	.14710	.00598	0.0
Year built:			
30 years ago or more	.00236	.00344	1.5
Not reported	.48828	.24611	0.5
Nativity:			
Native	.00024	.00052	2.2
Foreign	.00760	.03043	4.0
Residence 5 years ago:			
Same house	.00013	.00435	33.5
Different house, same county	.00351	.01041	3.0
Different county or abroad	.00385	.00571	1.5
Educational attainment:			
Highest grade attended, not completed	.01438	.00816	0.6
Elementary 1-2	.05840	.20191	3.5
Elementary 8	.00296	.01651	5.6
Grade 9 or more	.00030	.00077	2.6
High school 4	.00208	.00290	1.4
College 1	.00667	.01288	1.9
College 1 or higher	.00192	.00195	1.0
College 5 or higher	.01599	.00978	0.6
Not reported	.23341	.12620	0.5
School Enrollment:			
Kindergarten or first grade	.00122	.00727	6.0
Elementary 8	.00000	.02104	- [1]
High school 1	.01875	.01806	1.0
High school 4	.01865	.03161	1.7
College 1	.14344	.03681	0.3
College 5 or more	.44102	.08015	0.2
Public elementary	.00000	.00385	- [1]
Private elementary	.04042	.01673	0.4
Public high school	.00675	.01230	1.8
Private high school	.06761	.00000	0.0
Not reported	.58555	.26258	0.4
Number of children:			
None	.00274	.00398	1.5
1-3 children	.00040	.00064	1.6
3 or more children	.00081	.00112	1.4
5 or more children	.00548	.00322	0.6
Labor force:			
Unemployed	.07552	.04522	0.6
Worked less than 35 hours last week	.00281	.00385	1.4
Occupation groups:			
Professional, technical	.00026	.00000	0.0

[1] When 1960 estimate is 0.0, this ratio is not defined.

TABLE 1.--COMPARISON OF CORRELATED COMPONENT OF RESPONSE RELVARIANCES FOR IDENTICAL ITEMS: 1960 AND 1970 CENSUSES--CONTINUED

Characteristic	Response relvariances		Ratio of 1970 to 1960 relvariances
	1960 (1)	1970 (2)	(3)
Occupation groups--continued			
Farmers, farm managers	.00868	.25916	29.9
Clerical	.00247	.00381	1.5
Sales workers	.00000	.00766	- [1]
Craftsmen, foremen	.00408	.00000	0.0
Operatives	.00281	.00573	2.0
Farm laborers, paid workers	.04845	.00630	0.1
Wage and salary income:			
None	.00090	.01565	17.4
$2,500 or more	.00026	.00155	6.0
Males, less than $3,000	.00745	.01218	1.6
Females, less than $3,000	.00322	.00639	2.0
Males, $3,000 to $4,999	.00425	.00000	0.0
Females, $3,000 to $4,999	.01060	.00343	0.3
Males, $5,000 to $6,999	.00320	.00711	2.2
Females, $5,000 to $6,999	.02848	.00649	0.2
Males, $7,000 to $9,999	.01118	.00867	0.8
Females, $7,000 to $9,999	.00000	.00841	- [1]
Males, $10,000 or more	.00256	.00056	0.2
Females, $10,000 or more	.00000	.00000	0.0
Not reported	.19134	.08275	0.4
Self-employment income:			
$2,500 or more	.00704	.00000	0.0
Males, less than $3,000	.01235	.02391	1.9
Females, less than $3,000	.07884	.02163	0.3
Males, $3,000 to $4,999	.00000	.04469	- [1]
Females, $3,000 to $4,999	.16999	.00000	0.0
Males, $5,000 to $6,999	.04805	.13789	2.9
Females, $5,000 to $6,999	.38484	.00000	0.0
Males, $7,000 to $9,999	.00000	.00000	0.0
Females, $7,000 to $9,999	.00000	.20046	- [1]
Males, $10,000 or more	.02356	.03072	1.3
Females, $10,000 or more	.00000	.40936	- [1]
Not reported	.19721	.13332	0.7
Other income:			
$2,500 or more	.02868	.00511	0.2
Males, less than $3,000	.00003	.00378	126.0
Females, less than $3,000	.00709	.00178	0.3
Males, $3,000 to $4,999	.05200	.03701	0.7
Females, $3,000 to $4,999	.13785	.00000	0.0
Males, $5,000 to $6,999	.02590	.05832	2.3
Females, $5,000 to $6,999	.00000	.20822	- [1]
Males, $7,000 to $9,999	.00000	.08168	- [1]
Females, $7,000 to $9,999	1.18107	.00000	0.0
Males, $10,000 or more	.00000	.13260	- [1]
Females, $10,000 or more	2.44674	.24646	0.1
Not reported	.18532	.18390	1.0
Veteran status:			
World War II veterans	.00304	.00902	3.0

[1] When 1960 estimate is 0.0, this ratio is not defined.

TABLE 2.--1970 ESTIMATED CORRELATED RESPONSE RELVARIANCE AND ESTIMATED SAMPLING RELVARIANCES FOR SELECTED HOUSING ITEMS FOR AN ENUMERATION BY ONE ENUMERATOR IN AN AREA OF 2,500 HOUSING UNITS

Characteristic	Percent of housing units	Relvariances		Ratio of response to sampling variance
		Response	Sampling	
All housing units	100.0	-	-	-
Occupied units	94.9	.00017	.00040	.42
Vacant	5.1	.05866	.13841	.42
Vacant units,	100.0	-	-	-
Vacancy status:				
For rent	44.1	.06713	.19357	.35
For sale	14.8	.30926	.91600	.34
For rent or sale	58.9	.05678	.10550	.54
Other vacant	41.0	.00715	.21767	.54
Not reported	10.3	1.7419	1.3796	1.26
Occupied unit,	100.0	-	-	-
Tenure:				
Owned or being bought	63.5	.00001	.00309	.00
Cooperative, or condominum	1.4	.25748	.22754	1.13
Rented, cash rent	33.8	.00035	.01059	.03
Rented, no cash rent	1.4	.00506	.57512	.01
Not reported	1.4	.24887	.60020	.41
Vacant units,	100.0	-	-	-
Duration of vacancy:				
less than 2 months	43.6	.26728	.19225	1.39
2 to 6 months	31.1	1.0032	.33076	3.03
6 months or more	25.2	1.2087	.44361	2.72
Not reported	16.3	1.7457	.82702	2.11
All housing units,	100.0	-	-	-
Number of rooms:				
1 room	1.4	.62714	.41483	1.51
2 rooms	2.7	.08184	.24460	.33
3 rooms	10.9	.00000	.05127	.00
4 rooms	19.0	.00586	.02935	.20
5 rooms	24.7	.00292	.02246	.13
6 rooms	21.5	.00000	.02747	.00
7 rooms	10.8	.00049	.06269	.01
8 rooms	5.7	.00617	.12068	.05
9 rooms or more	3.2	.09479	.22003	.43
Not reported	1.5	.39478	.54493	.72
Occupied units,	100.0	-	-	-
Persons per room:				
.50 or less	48.4	.00000	.00841	.00
.51 to .75	25.1	.00929	.02480	.37
.76 to 1.00	19.9	.00365	.03242	.11
1.01 to 1.50	5.0	.00000	1.5741	.00
1.51 or more	1.5	.16008	.50754	.32
Occupied units,	100.0	-	-	-
Number of persons in unit:				
1 person	15.4	.00843	.04308	.20
2 persons	29.3	.00179	.02011	.09
3-4 persons	34.8	.00332	.01567	.21
5-6 persons	16.0	.00087	.04305	.02
7 persons or more	4.6	.00658	.17762	.04
All units,	100.0	-	-	-
Year built:				
1969 or 1970	3.5	.04061	.13863	.29
1965 to 1968	10.2	.00822	.04860	.17

TABLE 2.--1970 ESTIMATED CORRELATED RESPONSE RELVARIANCES AND ESTIMATED SAMPLING RELVARIANCES FOR SELECTED HOUSING ITEMS FOR AN ENUMERATION BY ONE ENUMERATOR IN AN AREA OF 2,500 HOUSING UNITS--CONTINUED

Characteristic	Percent of housing units	Relvariances		Ratio of response to sampling variance
		Response	Sampling	
Year built--continued				
1960 to 1964	15.4	.00066	.03135	.02
1950 to 1959	27.1	.00085	.01483	.06
1940 to 1949	14.6	.01559	.03774	.41
1939 or earlier	29.3	.00343	.01109	.31
Not reported	6.9	.24506	.10600	2.31
Occupied, owned units and vacant units for sale	100.0	-	-	-
Value:				
Less than $5,000	1.6	.00000	.94179	.00
$5,000 to $9,999	9.0	.00000	.12638	.00
$10,000 to $14,999	17.9	.00000	.06041	.00
$15,000 to $19,999	23.6	.00000	.04358	.00
$20,000 to $24,999	17.7	.00852	.06463	.13
$25,000 to $34,999	18.6	.00215	.05355	.04
$35,000 to $49,999	8.2	.04484	.13612	.33
$50,000 or more	3.3	.04745	.29936	.16
Not reported	2.4	.17789	.62656	.28
Occupied, rented units,	100.0	-	-	-
Gross rent:				
No cash rent	4.2	.00820	.55334	.01
$1 to $29	.1	.00000	3.5817	.00
$30 to $39	2.0	.00000	1.2606	.00
$40 to $49	1.6	.19743	1.4067	.14
$50 to $59	2.6	.05732	.89119	.06
$60 to $69	4.2	.02585	.57597	.04
$70 to $79	5.3	.00000	.43266	.00
$80 to $99	12.8	.03343	.15896	.21
$100 to $119	14.0	.02135	.14496	.15
$120 to $149	19.7	.00000	.08882	.00
$150 to $199	24.2	.05533	.05936	.93
$200 to $249	6.3	.03228	.32178	.10
$250 to $299	2.2	.27808	.94106	.30
$300 or more	1.5	.13675	1.5607	.09
Occupied units,	100.0	-	-	-
Type of family:				
Husband-wife	72.5	.00030	.00295	.10
Other, male head	2.2	.03961	.37672	.11
Other, female head	8.3	.00366	.09425	.04
Male primary individual	5.7	.00581	.13653	.04
Female primary individual	11.3	.00900	.06352	.14

TABLE 3.--1970 ESTIMATED CORRELATED RESPONSE RELVARIANCES AND ESTIMATED SAMPLING RELVARIANCES FOR SELECTED POPULATION ITEMS FOR AN ENUMERATION BY ONE ENUMERATOR IN AN AREA OF 7,500 PERSONS

Characteristic	Percent of people	Relvariances		Ratio of response to sampling variance
		Response	Sampling	
All persons,	100.0	-	-	-
Sex:				
Male	48.2	.00019	.00165	.11
Female	51.8	.00016	.00144	.11
Not reported	1.8	.13087	.43902	.30
All persons,	100.0	-	-	-
Race:				
White	90.8	.00019	.00082	.23
Negro	8.2	.02494	.08893	.28
Other	.9	.00000	1.1032	.00
NA	2.0	.21526	.45120	.48
All persons,	100.0	-	-	-
Age:				
0 to 14 years	29.7	.00038	.00813	.05
15 to 24 years	15.8	.00000	.01821	.00
25 to 34 years	12.7	.00077	.02350	.03
35 to 44 years	12.0	.00010	.12171	.00
45 to 54 years	12.1	.00285	.02688	.11
55 to 64 years	8.9	.00553	.04052	.14
65 years and over	8.8	.00397	.04132	.10
Not reported	4.1	.27619	.17607	1.57
All persons,	100.0	-	-	-
Marital status:				
Married	65.4	.00000	.00241	.00
Widowed	7.1	.00000	.05209	.00
Divorced or separated	4.7	.00648	.08467	.08
Never married	22.8	.00069	.01442	.05
Not reported	2.8	.25637	.26276	.98
Persons 5 years and over,	100.0	-	-	-
Residence in 1965 :				
Same house	54.0	.00056	.01108	.05
Different house, same county	23.3	.00420	.04090	.10
Different county, same State	8.3	.01418	.12929	.11
Different State	8.2	.01457	.13410	.11
Abroad	1.3	.10623	.92275	.12
Moved, residence in 1965 not reported	3.7	.13124	.21773	.60
State reported, but place or county not reported	1.1	.09470	.58595	.16
Not reported	3.8	.09899	.21210	.47
Persons 3 to 34 years, attending school	100.0	-	-	-
School enrollment:				
Nursery school	13.3	.02650	.60650	.04
Kindergarten or elementary 1	43.3	.01054	.07930	.13
Elementary 2-7	43.3	.00257	.01692	.15
Elementary 8	6.8	.01408	.15117	.10
High school 1	6.7	.00235	.15919	.01
High school 2-3	12.6	.00000	.08225	.00
High school 4	5.7	.02875	.20418	.14
College 1-4	8.1	.01728	.17529	.10
College 5 or more	1.4	.06412	.99288	.06
Not reported	3.1	.24390	.68511	.36

TABLE 3.--1970 ESTIMATED CORRELATED RESPONSE RELVARIANCES AND ESTIMATED SAMPLING RELVARIANCES FOR SELECTED POPULATION ITEMS FOR AN ENUMERATION BY ONE ENUMERATOR IN AN AREA OF 7,500 PERSONS --CONTINUED

Characteristic	Percent of people	Relvariances		Ratio of response to sampling variance
		Response	Sampling	
Persons 25 years and over,	100.0	-	-	-
Educational attainment:				
Never attended, nursery school or kindergarten	1.3	.24222	.41077	.59
Elementary 1 to 4	2.5	.04238	.21547	.20
Elementary 5 to 7	7.9	.01590	.06649	.24
Elementary 8	11.1	.01565	.04559	.34
High school 1 to 3	19.6	.00731	.02300	.32
High school 4	34.7	.00245	.01097	.22
College 1 to 3	11.4	.00414	.04245	.10
College 4	6.5	.00123	.07906	.02
College 5 or higher	4.9	.00704	.10915	.06
Not reported	7.9	.12620	.08400	1.50
Persons 14 and over,	100.0	-	-	-
Employment status:				
In labor force	57.7	.00021	.00241	.09
At work	52.6	.00026	.00304	.08
With a job, not at work	1.7	.02551	.24169	.11
Unemployed	2.4	.04543	.16716	.27
Armed Forces	1.0	.06091	.35738	.17
Not in labor force	42.3	.00039	.00449	.09
Not reported	4.6	.28882	.12940	2.23
Persons 14 and over who worked since 1960,	100.0	-	-	-
Industry:				
Agriculture, forestry and fisheries	2.2	.04563	.25259	.18
Mining	0.3	.36765	1.8685	.20
Construction	5.2	.00000	.10173	.00
Manufacturing, durables	15.8	.00000	.02786	.00
Manufacturing, non-durables	10.5	.00000	.04591	.00
Transportation, communication and other public utilities	6.2	.00501	.08264	.06
Wholesale trade	4.3	.01883	.12099	.16
Retail trade	18.5	.00109	.02523	.04
Finance, insurance and real estate	6.2	.01322	.08062	.16
Business and repair services	3.3	.02189	.16517	.13
Personal services	4.3	.01476	.12685	.12
Entertainment and recreation	1.0	.06610	.56488	.12
Professional services	16.5	.00649	.02919	.22
Public administration	5.7	.00716	.089047	.08
Not reported	6.8	.19010	.09283	2.05
Persons 14 and over who worked since 1960,	100.0	-	-	-
Occupation:				
Professional, technical, and kindred workers	14.2	.00000	.03484	.00
Managers	7.3	.01866	.06381	.29
Sales workers	9.1	.00549	.05179	.11
Clerical	22.4	.00331	.01658	.20
Craftsmen	12.7	.00111	.03326	.03
Operatives	12.8	.00560	.036604	.15
Transport equipment operators	3.0	.01640	.16553	.10
Laborers, except farm	3.9	.01440	.13061	.11
Farmers	.4	.26662	1.3336	.20
Farm laborers	1.3	.00000	.46027	.00
Service workers	11.4	.00353	.04242	.08
Private household workers	1.5	.01708	.34988	.05
Not reported	7.8	.11685	.07733	1.51

WORKLOAD AND CORONARY HEART DISEASE

Sidney Cobb, Butler Hospital and Brown University

Before starting out to explore the relationship between workload and coronary heart disease, it is well to define terms and specify the model under consideration. The model is presented in Fig. 1. Starting at the right hand end, we are concerned with coronary disease which is basically atherosclerosis of the coronary arteries. The atherosclerotic process narrows the arteries to the point at which the blood flow to the heart muscle is insufficient. This causes pain known as angina pectoris. If the artery is occluded myocardial infarction with damage to the heart muscle may ensue. Moving to the left in the diagram, there are listed a series of variables mostly physiological that are known to enhance the risk for coronary disease. The evidence for increased risk associated with high levels on these variables is well documented and in most instances replicated. For present purposes, I propose that we accept these associations as given, and address ourselves to the question as to whether these risk factors are influenced by workload. Before going on to that, we should take note of the fact that these risk factors are by no means fully independent of each other. Not only is it known that blood sugar is correlated with catecholamine output and serum cortisol levels but a good deal is known about the mechanisms involved. Other such associations are known among these variables but again we should only have to go into this in detail if we were to find ourselves in a position to write a prediction equation. Unhappily, we are no where near the stage at which that could be done with any validity. The arrows imply a direction of causation but must be accepted with caution because we don't always have good evidence of the direction of causation, if any, in the associations observed. An arrow to an arrow implies an interaction in the statistical sense.

Now we come to workload. On the face of it, one might think this to be a nice simple variable. Unfortunately, this is not the case. We do both mental and physical work. Here we are concerned only with mental work. As a matter of fact, physical work may have some life saving aspects in the event of a coronary occlusion because those who exercise regularly have a more extensive network of blood vessels to the heart muscle so a single obstruction is less likely to be fatal.

Workload has both quantitative and qualitative aspects and these are readily separable as demonstrated by (French et al 1965). They found in a study of University faculty and administrators that among professors low self esteem was significantly related to qualitative overload but not to quantitative overload. Quite the reverse was true of the administrators who thought ill of themselves if they were not keeping up with the work but who had no reason to suppose that they were technically qualified to deal with all the problems that they face. "That's what consultants are for." There is little evidence that qualitative overload is associated with coronary disease, so we are concerned with quantitative overload.

Finally quantitative mental work load may be just right for the individual or it may be too little or too much. This is a function of the ability of the individual. What is underload for one man may be overload for another. Work underload has been related to a variety of things from Adam Smith (1776) to the present day but the evidence is meager that it is related to coronary disease. We are then left focussing on quantitative overload of mental work. By and large, I will be reporting studies involving subjective assessment of the overload. It should of course be clear that as pointed out first by Parkinson (1957) and later by Sales (1969) the subjective assessment of what constitutes a full workload may not coincide with that of independent observers.

There are a variety of social variables that contribute to workload. First Kahn et al (1964) have pointed out that those persons who make demands on one are much more apt to be in conflict over priorities, thereby generating overload, than they are to be in direct conflict as to what one should do. Similarly the consequence of role ambiguity is overload for those who in the face of uncertainty try to cover all the fronts. Finally what Terreberry (1968) has called complexification and Lipowski (1973) has referred to as a surfeit of attractive information inputs, also generates a subjective sense of overload of mental work.

It is now appropriate to ask what is the evidence connecting this kind of overload with coronary disease. The evidence is muddy but certainly quite voluminous. The mud is contributed by a lack of dimensional clarity. Factors of responsibility and emotional arousal are all interwoven in a way which makes it quite

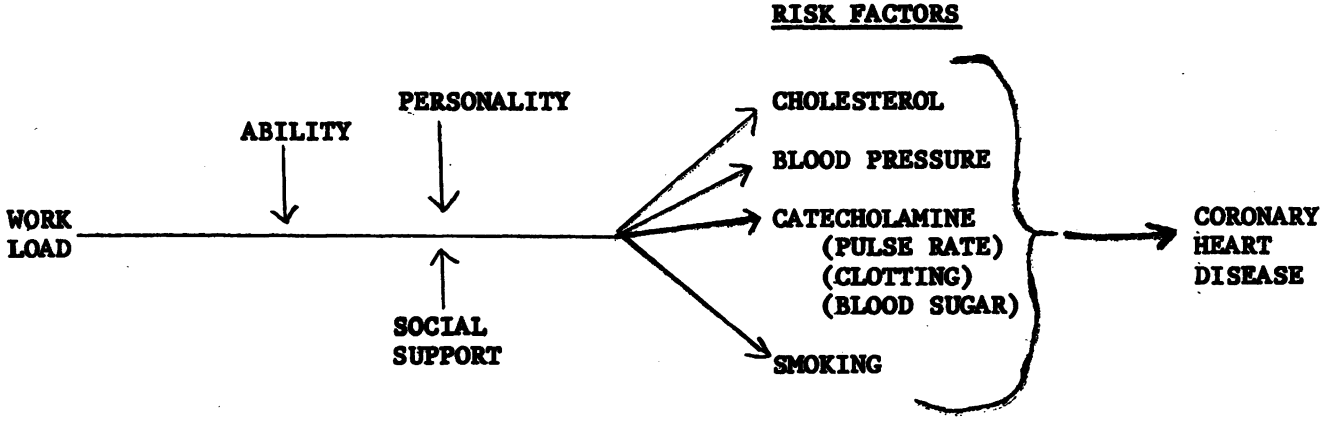


Fig. 1 A model of the possible relationship of workload to coronary heart disease.

possible to interpret these studies in several different ways. Now is not the moment to list these studies and criticize them one by one. Suffice it to say that there is evidence to suggest that work overload particularly among those who are conscientious, ego involved and persistent may be associated with an excess frequency of coronary heart disease.

The remainder of my time will be devoted to an examination of the evidence for an association between quantitative overload of mental work and the risk factors in coronary heart disease. I will make no effort to be all inclusive in the review, for the state of the art is still such that there are many ways to fall short of finding an association that truly exists. Most of these are in the area of measurement of the relevant psycho-social variable. I will therefore emphasize the positive.

For this presentation, I will emphasize cholesterol, blood pressure, catecholamine (adrenaline and nor-adrenaline) output and smoking behavior. Looking first at cholesterol, we go back to the original report of Friedman, Rosenman and Carroll (1958) on tax accountants. It was demonstrated that cholesterol levels rose as tax deadlines approached and fell afterward. Since then, there have been five studies of medical students at examination time each showing a significant increase. Across all individuals in the five studies, the average increase from usual level to exam level is 31 mg%. This is a difference, which if maintained, would increase the probability of coronary disease by about 30%. Similarly Caplan & French (1968) using both objective and subjective workload measures found associations with concurrent cholesterol levels. Likewise Sales (1969) using an experimental set up to produce overload of mental work with student subjects found a rise in cholesterol in the course of a one hour period that just fell short of significance. I mention this study because it is the one study in which we can be reasonably sure that the independent variable isn't laced with other factors such as negative affect and responsibility for persons.

There seems to be ample evidence that catecholamine output is related to work load. Frankenhaeuser (1971) has shown that mental work increases the output of both adrenalin and nor-adrenalin. This is true whether the mental work is a boring job requiring continuous vigilance or a complex and difficult time paced task. As might be expected, the difficult task increases the output more than the boring task. Similarly Levi (1972) has shown that under an experimental piecework scheme that led to doubling the output of invoicing clerks, the urinary excretion of adrenaline and nor-adrenaline were consistently increased on the days when piecework was the basis of pay.

Since it is known that increased catecholamine output is associated with increased pulse rate, increased blood sugar and increased stickiness of the platelets therefore increase probability of clots forming in the blood vessels and each of these in turn are reasonably well associated with coronary heart disease, we have a pretty strong chain of associations. Incidentally, those of you who are coffee drinkers will be interested to know that coffee interacts with environmental stress. That is, it produces elevation of the catecholamine output only in the presence of stress. (Cobb & Rose, 1973)

The evidence for the association of elevated blood pressure with quantitative overload of mental work is limited but quite striking. Of all the people I have ever seen doing mental work, the air traffic controllers reach the highest levels of rapid accurate performance. These levels of performance are demanded only at intervals. A given controller is not likely to have more than one episode per day of maximum arousal except in bad weather. Though the peaks of workload surmounted by controllers are higher than any that most people can imagine, it is not clear how much these should be seen as overload. Clearly, the men mostly enjoy this work and have selected themselves into this job. However, if the principle dimension of stress contributing to the strain they experience is really overload, then the price they pay is high, for hypertensive disease is about four times as common among air traffic controllers as it is among the fliers who are subjected to the same physical evaluation at the same intervals by the same aeromedical examiners (Cobb & Rose, 1973). The probability that this relationship is in fact an etiologic one is increased by the fact that adjusting for age, the frequency of hypertension is greater at airports with high traffic density than it is at airports with low traffic density.

Caplan (1971) has shown an interesting set of interactions. He has demonstrated that the association of workload and diastolic blood pressure is significantly positive only among those who are either flexible, persistent or are unsupported by their subordinates.

Neither of these studies are free of contamination with other variables. Responsibility is particularly likely to account for part of the findings, However, until we get studies in which work load and responsibility are measured separately and concurrently, we won't be able to draw clear conclusions.

With regard to smoking Caplan (1971) has produced unequivical evidence that no matter what measure of overload he used, NASA personnel who smoked had higher workloads than those who were ex-smokers or who never smoked. A direction of causation is implied by the fact that the quit rate, i.e. exsmokers divided by all who ever smoked, is highest in those with low workloads.

In conclusion, I feel that I have made a moderately good case for quantitative overload of mental work contributing to coronary heart disease via several paths. The importance of this report lies not in the substantive conclusion which still needs more work but rather in the demonstration that we have come a long way in psychosocial epidemiology in the last dozen years. When I first went to the Institute for Social Research in 1961, my colleagues and I were looking at the association of two global variables social status and dispensary visits. Now I am able to present to you a substantial body of evidence relating a single well defined social variable, quantitative overload of mental work, to a single disease entity, coronary disease, via four risk factors, cholesterol, blood pressure catecholamine levels and smoking behavior. In view of our progress in concept formation and measurement, I think some major advances in psychosocial epidemiology are likely in the next ten years.

REFERENCES

Caplan, R. D. Organizational Stress and Individual Strain: A Social-Psychological Study of Risk Factors in Coronary Heart Disease Among Administrators, Engineers and Scientists. Doctoral dissertation, Univ. of Mich., 1971.

Caplan, R. D. & French, J. R. P. Jr. Physiological Responses to Work Load. Unpublished Manuscript Institute for Social Research, 1968.

Cobb, S. & Rose, R. M. Hypertension, Peptic Ulcer and Diabetes in Air Traffic Controllers. JAMA 224: 489-492, April 23, 1973

Frankenhaeuser, M. Experimental Approaches to the Study of Human Behavior as Related to Neuro-endocrine Functions. Chap. 4 in Levi, L. (Edit) Society Stress and Disease I The Psychosocial Environment and Psychosomatic Diseases. London Oxford Univ. Press 1971

French, J.R.P. Jr., Tupper, C.J. & Mueller, E. F. Work Load of University Professors. Cooperative Research Project No. 2171. U.S. Office of Education, Ann Arbor, Univ. of Mich. 1965

Friedman, M Rosenmann, R.H. & Carroll, V. Changes in the Serum Cholesterol and Blood Clotting Time in Men subjected to Cyclic Variation in Occupational Stress, Circulation 18: 852-861, 1958.

Levi, L. Stress and Distress in Response to Psychosocial Stimuli. International Series of Monographs in Experimental Psychology, 17. Oxford, Pergamon Press, 1972, pp. 106-118.

Lipowski, Z.J. Affluence Information. Inputs and Health - Soc. Sci. & Med. 7: 517-529, 1973.

Kahn, R.L. Wolfe, D.M. Quinn, R.P. Snoek, J.D. & Rosenthal, R.A. Organizational Stress: Studies in Role Conflict & Ambiquity. New York, Wiley, 1964.

Parkinson, C.N. Parkinson's Law and Other Studies in Administration. Boston Houghton, Mifflin, 1957.

Sales, S.M. Differences Among Individuals in Affective, Behavioral, Biochemical and Physiological Responses to Variations in Work Load. Doctoral Dissertation, Univ. of Mich., 1969.

Smith, A. The Education of the Worker. 1776. Op. Lit. by A. O. Lewis, Edit Of Men and Machines. New York Dutton, 1963

Terreberry, S. The Organization of Environments. Doctoral Dissertation, Univ. of Michigan, 1968.

POLICY RESEARCH AND HEALTH STATUS

Selma J. Mushkin, Public Services Laboratory, Georgetown University

What policy research concepts have an impact on the health of the population? The question may surprise you. Why consider policy research as a means of improving health care? More specifically, why, in the context of a Carnegie-financed inquiry on nonhealth factors that make a difference in health status, should policy research as such be identified?

Definitions and Some Implications

Policy research is defined as research that uses analytical and evaluative techniques to examine public policies, programs, and projects. It is essentially application of scientific method, with a continuing questioning of problem, process, and research design and the probing of the findings for their significance.

Analysis involves examination of many questions: What are the purposes or objectives of the policy or program? How can they be measured so that we can quantify and record success in achieving them? What are the optional methods of attaining the objectives sought? What are the comparative costs of each method--now and later? What are the gains that may be expected--immediately and in the longer run? And in any case, how certain are the estimates; how sensitive are they to the assumptions on which they are built?

Clearly the drive toward policy analysis and evaluation is not restricted to health programs. Nonhealth programs with health impacts also come into question as part of the inquiry into program outputs for resources spent. What types of nonhealth activities are at issue to which policy research can be directed? In various combinations the factors are these:

Consumption and "problem item with some health impact"	Illustrative public policies
nutrition and food safety	food and drug laws; school lunch; food stamps
housing	public housing; rental allowances
safety of property and person	fire and police protection; auto regulation; driver regulation; alcoholism control; street lighting
transportation	highway and street construction; air-transport safety controls
income security	income maintenance
job opportunities and advance	work training; employment services; public service employment
leisure time activities	parks; playgrounds; ball fields; libraries
education	schooling; right to read; compensatory education.

Benefit assessments underscore the commonality of program purposes. Many public programs involve pricing human life, programs that are not directly associated with health care and that do not bear the label of health programs. The most familiar example is the highway construction program, where life-saving costs or death and disability costs must be equated in deciding whether or not to construct a clover leaf or add another highway lane. (1,2) Airway safety and convenience in time may sometimes be traded off in the congested air terminals of the major cities. (3,4) Building codes clearly have health and safety purposes, as do water supply and waste-treatment measures. (5,6) Energy production not only affects jobs and industrial development but also is linked to hazards of atomic wastes and water and air pollution. (7) For many programs or components it is difficult to draw a line between health and nonhealth programming.

Three Health and Nonhealth Programs

The broad questions are far from new, but it is important now to restate the issues of policy research as a health status input and to consider it in the specific terms of program analysis. Three policy issues are summarized here: lead poisoning prevention in children, transportation as a component of health care, and drug addiction and abuse. (8,9,10,11) A series of inquiries would help answer the question: Would policy research make a difference in health status? Among the issues that need to be considered are those ennumerated below. Only a start to an examination of the issues is made in this paper.

- Are the program options designed specifically to meet health objectives or is the impact on health a by-product.

- If health purposes are involved, are there special policy characteristics of the health care?

- For the nonhealth programs, what are the health data and research findings and what additional knowledge do we need for policy purposes? And if we had those data and that knowledge, would the health status of the people be improved.

Lead Poisoning Prevention in Children

Lead poisoning in children is typically regarded as an environmental health problem to be prevented by removing the hazard--in this instance the lead paint used in old housing. Old housing can be torn down, or leaded surfaces removed by various methods.

Analysis of lead poisoning in children in the specific environment of Washington, D.C., raised first the question about statistical reporting of health problems. National data were not available when the study was initiated. The size of the lead risk was defined by prevalence counts obtained by screening in New York City where, in an initial count, it was reported that 25 percent of the children aged 1-6 living in old homes had "undue" blood-lead levels. Blood-lead level defined as "undue" in the standard adopted by the Surgeon General was 40 micrograms of lead per 100 milliliters of whole blood.

Analysis suggested that young children who were teething and mouthing objects had a higher rate. As children grow older the lead is excreted and the blood lead levels are lowered. But does the subsequent excretion correct the damage done? The answer from those physicians expert in lead poisoning was "no," a "no" that was followed by the suggestion that the peak rate cited for the young ages was the minimum exposure to the hazards from lead. What was the actual risk? When corrected for the observed (but poorly understood) higher lead levels in the warm summer months and also for the few children exposed after the first three years of life, the proportion of the children living in hazardous neighborhoods and exposed to an undue blood-lead level before reaching school age was estimated for the District to reach perhaps as high as 50 percent in some neighborhoods. Thus analysis showed first the inadequacies of the usual statistical counts and then opened the question of what is in fact medically known about the damage lead poisoning does to the child.

But the policy research raised still added questions: housing repair practices called for inspection of dwellings in which the children with elevated blood-lead levels lived. An order for a cleanup was issued when inspection showed lead content above the standard in painted surfaces of the dwellings. But what is to be the environment of the child while the repairs are being made. In New York City, children were reported to have been hospitalized for 30 days at a cost in 1971-1972 of about $100 a day, or roughly $3,000 per child. In Washington, D.C., consideration is being given to halfway houses where the children can stay while undergoing outpatient chelation treatment. Differences between outpatient and inpatient treatments are roughly $150-$500 per case, even when the length of hospital stay is only the five days required for the chelation.

But there is a more specific health program choice that is pertinent in the policy research--namely, the practice of the physicians and the medical community generally to place resources at the disposal of the "interesting case" to the neglect of the more parochial and perhaps more preventive treatments.

The child in a coma with encephalepathy receives much care and at great expense, so great in New York City, that the city opted for higher standards for undue blood-lead levels before cautionary action and re-screening of children generally. Much in the way of resources is devoted to the care of those who are brain damaged (35 percent of the cases) despite the care. At the other extreme, where preventive health care is possible for far larger numbers at only a small screening cost, with blood-sample testing and treatment, the access to such care is restricted to children with 60 micrograms of lead per 100 milliliters whole blood to reduce outlays. Medical practices thus come into question. New action is called for involving physicians' training, a better understanding of the risks of disablement, and the chances of getting well, with and without subsequent impairment of varying degrees.

Preventive action in lead poisoning is usually viewed, as indicated at the outset, not as a health program but as a cleanup of housing to remove lead hazards. For Washington, D.C., where 47 percent of the houses were built at a time when lead was used in paint, costs of total housing cleanup would be high. They would be high in direct costs, and perhaps higher still in the incentive that such requirements would give to property owners to demolish the houses in favor of lucrative parking lots and thus to reduce the supply of housing, especially for poor families with many young children.

I estimated the cost of a full cleanup in the District of $69-$138 million. And

this is a gross underestimate even when account is taken of the newer areas of the city, presumably free of lead hazard, where many young children live.

What then?

Careful review of the problem suggested (1) that the only feasible low-cost option was a program of parental education, and (2) that the families hardest to reach were perhaps those on welfare, making the program policies more nearly a welfare agency issue than one of the housing or health agency.

What does policy research tell us about means of communication with the poor, with the nonreading public, and even with those who are not functionally literate? Dr. Henrietta Sachs, of Chicago, reported X-ray diagnosis in lead cases. When she was asked about that practice, Dr. Sachs answered by asking in turn: How else is it possible to make sure that parents know about the disease in their children when there are no apparent symptoms? (One aspect of the disease is its asymptomatic characteristic: one study found that, in 53 percent of the cases where blood-lead levels were 2-3 times above the "undue" standard, there were few if any symptoms.)

The question of communication may be posed in the dreary terms of a specific case of children who live with a drug addict parent in a boarded-up house marked "condemned." If policy research is to have an impact on health, perhaps we must first understand how to reach those who must be reached, and then find the methods for inducing behavior change without intruding on the dignity of the individual or his personal rights.

There are many more lessons that the lead-poisoning case had for health programs: the need for research on disease detection methods in laboratories, new methods of screening with a micro blood sample only, and new methods of treatment that would be more specific to the metal "lead." New engineering research appeared to be indicated on methods of detecting lead in housing surfaces that would be less costly, require fewer inspectors, and perhaps require less conversion of statutory standards on "unsafe" housing. And at the behavioral science end, there were posed familiar questions about users of screening programs. Why did the children disappear when the mobile clinic came to the neighborhood? To what kinds of health personnel are the families in the ghetto most responsive--the neighborhood aide, the professional physician, the public health nurse, the home health aide?

Transportation as a Component of Health Care

The District of Columbia runs three major transportation services having a special bearing on health care: a contractual service for transportation of vocational rehabilitation patients who cannot use public transportation, a public ambulance service for emergency transportation run by the Fire Department, and a small transportation service run by the D.C. Department of Human Resources for those too old, too infirm, and too poor to use public transportation. (10)

The initial objectives of the transportation study were simply to reduce the health hazards to persons who do not show for treatment or screening and to reduce the loss in time for physicians who scheduled appointments for those persons. As the study progressed--and it still remains incomplete--the health objectives shifted in part toward reducing the costs of health care by substituting transportation services for new hospital facilities, including new beds at Cafritz Hospital and a new hospital facility in the Northeastern part of the city. Important nonhealth objectives came to the fore. Those objectives, in brief, were maintaining income and jobs for low- and moderate-income persons in the city and reducing loss of earnings and of employment time for those with disabled persons in the home or with children who required emergency backup or occasional assistance in reaching medical attention.

Two health care practices are bound up in transportation policy research. The first is the trend toward ambulatory primary care that provides easier access to care for patients; the second is the trend toward lower cost methods. Together, they suggest more transportation services to health facilities, with escort services and links to caregiver services, as well as differentiated transportation services that reflect differences in individual needs and costs.

Many factors over the years have made for ever-increasing use of hospitals as the center of health services. Hospitals are known as centers of specialities and of quality health care. The transition to the hospital center has had a profound effect, as the hospital has built up its emergency room, social service departments, ambulatory clinics, liaisons with other medical care facilities such as extended care facilities, and home health care units. Through the establishment of neigborhood centers the hospital has extended its interests actively into the community, both to keep costs down and to reach persons who otherwise would not be reached. And more recently, hospitals have been working toward the development of health maintenance organizations that will permit medical care to

be provided on a prepayment basis, with emphasis on preventive care.

Patient and family use of hospital centers for outpatient care should come as no surprise. Hospital facilities are manned around the clock, while physicians, especially in this age of specialists, are hard to reach; care in hospitals is known to be good; care outside hospitals is often not accessible. Emergency rooms in hospitals require no appointments; in the out-of-hospital medical world there are long waits for appointments, and the hours in which appointments will be made are limited. Home visits are rare today, and even health maintenance organizations refer patients to emergency rooms in the late night hours. Heavy patient loads, the transient population characteristic of the district, and the extent to which the black population customarily uses clinics compound to result in extensive use of hospital outpatient services.

Increases in the use of high-cost ambulance services accompany the use of hospitals for outpatient services. In the District the ambulances run by the Fire Department are known popularly as "red taxis," which suggests the undue use of ambulances for nonemergency purposes and at no charge to the user. In New York City a similar use of ambulances and the long waiting time that resulted, on analysis, pointed to the application of screening devices to reduce use and to decentralization of dispatching points. The screening of calls is reported to have reduced wait times, to have improved the service from the perspective of the users, and to have reduced costs. Manned by trained nurses, the screening operation has, at the same time, provided access to medical advice and guidance for some persons, a service otherwise lacking in New York City--and also lacking in Washington. (11) From other cities including Baltimore, come cautions about the difficulties of over-the-telephone screening of patients. (12)

.Thus far, assessment of the transportation problems in the District of Columbia points to a trial at least of differentiated services that would provide people with easier access to health services and facilities or with home health care as indicated by screening arrangements. Among the services would be the dispatching of nonemergency vehicles that would be manned optionally by drivers, by drivers and attendants, or by drivers and caregivers, or ambulances would be sent in other cases, and in still others caregivers would be sent. Escort services also might be provided for children or aged persons, using public transportation.

When transportation to existing health care is inadequate and at the same time a shortage of facilities exists in any one region of a city, new hospital beds and outpatient facilities come to be demanded. Because of population shifts in the District of Columbia, a sizable segment of the population resides in areas where few health facilities are located. While the District as a whole has perhaps more than enough beds and other health facilities, in one area the bed-occupancy rates are running at 94 percent of capacity and the number of health practitioners is relatively low. With adequate transportation, available on demand, and with fees for the service scaled to cover age under Medicaid and other third party payments, building of new facilities could be avoided and perhaps lower cost neighborhood health services or group-practice units could be established as required.

The second objective relates to transportation services as a means to job continuity and income maintenance. This non-health objective still remains to be assessed. Mothers who work and have young children at home--no matter what the income level--clearly need the safeguard and assurance of quick response in an emergency. The same backstopping of a high-quality emergency care is also a requisite for those, both men and women, who have aged or chronically ill persons in the home. If emergency facilities are available promptly in response to calls, less emphasis has to be given to the ability of those watching over the children or the ill at home. One question is often uppermost: Can those at home, a grandmother for example, respond to emergencies? The question becomes less important when quick response service is available in a community.

Those providing home health services in the District of Columbia also called attention to the importance of transportation services of a less urgent kind for their patients. Such service is needed, they report to transport patients to medical and dental appointments, for consultants and observations, to get lab tests taken, and have braces and other devices fitted. (13)

What we have to determine in further analysis is the cost in lost time and earnings of the family member balanced against the cost of providing emergency services and home health aides as a contingency service, when needed. With another type of link between home health care and transportation services, it might be easier to transport young children to the clinic for working mothers, or those with large families, or to have someone stay with the children who need watching while the mother goes to the clinic herself or with another child. For many persons, particularly the poor, the availability of such services could make the difference

between job and no job, between independence and public dependency. The extent of the problem, the costs involved in provision of services, and the gains from that provision as an income- and job-maintenance method remain to be examined.

Drug Addiction and Abuse

The third case used for illustration is that of drug addiction, or more specifically, drug abuse and preventive measures. In a study that the Public Services Laboratory staff made for the National Commission on Marihuana and Drug Abuse, evaluations of the federal programs were assessed. (14) The starting point was a formulation of the objectives of a heroin addiction preventive program. Though health care is among the public objectives, other purposes are also prominent. The most apparent is the reduction of crime. A listing of objectives in drug control and drug prevention programs follows. It is presented partly because it is so suggestive of the multiple purposes of government programs, multiple purposes that further indicate the difficulties of gaining simplicity in analysis or evaluation. (15)

(1) To decrease the amount of drug addiction-related crime.

(2) To lower the number of drug addiction-related deaths.

(3) To rehabilitate the drug addict as a "useful" citizen within the community (employability).

(4) To decrease the number of new drug addicts.

(5) To enforce the laws pertaining to the flow of hard drugs and reduce the number of drug addicts.

(6) To reduce the number of drug addiction-related illnesses.

(7) To minimize the transfer of antisocial attitudes to other areas (crime, alcohol, other drugs).

The health care objectives of drug abuse prevention programs are fairly clear. They center first on the need to prevent a spreading of the disease. Each addict, it has been estimated, creates 20 addicts as he resorts to traffic in the drug to maintain his own habit. (14) And in an analytic vein a differentiation may be made between the extent of contagion in a community of young persons unaccustomed to addiction and in a community where addiction has long been a part of the scene. Reason suggests that the potential for contagion is greater in the former type of community than in the latter.

For each individual addict the objective in concept is a cure--freedom from addiction. A range of therapies has been developed to treat the addiction. Some of the treatment modalities are mental health methods, many call on the support of a therapeutic community to keep a former addict off the drug, and still others are chemotherapeutic.

Drug addiction prevention is one of a number of health measures in which the outcome of the resources spent may be favorable for neither the body politic nor the individual. For years, the Public Health Service provided care to addicts at its Lexington hospital. When the outcomes of that care were evaluated, the findings were simple: those who came for treatment continued to be addicts after treatment once they were outside the hospital again. (16) Similarly, few real cures over a sustained period are reported for the many different treatment modalities. It is not that temporary spells of nonaddiction are not achieved. And enthusiasm runs high among those who direct the treatment communities. This is a necessary part of the therapy, the belief of those who run the therapeutic communities in the efficacy of cure--the favorable outcomes. But few hard evaluations have been independently made that can throw light on a process that has a reasonable chance of long-term cures.

Does the temporary nature of the successes mean no further "tries?" On the contrary, in the case of drug addiction as in other social ills of deep concern to society, resources continue to be spent. If anything, there appears to be an even harder effort to find ways to meet some of the many objectives. But perhaps the suspicion is accurate that the spending policy, despite lack of success, originates in other than health purposes.

The high cost of crime associated with drug addiction has repeatedly been documented. The analysis starts with the size of the dosage required to support a habit of addiction, the cost of that dose, the required stealing to meet the cost, and therefore the amount of crime that is involved to support heroin addiction. Out of this analysis plainly comes some quantification of the size of the thefts. The total amounts stolen in support of the habit must exceed by sizable sums the amount of the heroin market. Nor can the heroin trading be neglected in the program assessment and design of options, for when the options are generated the behavior of the Mafia still remains to be factored. To reduce deaths from heroin and control the spread, the disease requires much analysis of nonhealth factors, including in this instance, crime.

Research on Policy Research Application

But does the research on specific questions find its way into public policy so that it can impact on health status? Or, more directly, are the policy research studies used? To have an impact, the studies clearly must be applied in policy or program action.

There are two schools of thought about the issue of use. The first alleges that the findings are not now used; (17,18) the second that the major research studies become a guide to policy decision. (19) The current Presidential budget gives evidence of the Administration's intent to use program evaluations in deciding which programs to continue, which to terminate, and which to expand. (20)

The processes of converting research into implemented policy have received altogether too little study and research support. There are many and not unimportant unknowns. A brief list of items for research as at least a partial agenda for research would include:

--- Research transmission channels--Some research fairly quickly enters into the debate. The Brookings Institution studies are among those that come into use with little delay. The Agenda volume prepared at the outset of the new Administration in 1968 (21) and the subsequent volumes by Schultze, Fried, Rivlin, and Teeters on national priorities are good examples of materials that are read by those who make decisions. (22,23,24) The summarizations of research included in those volumes stand up well, through their readability and balance, as illustrations of "means to communicate."

But what of research done in remoter places in the United States, and research that does not bear the endorsement of a prestigious institution? How is the work of researchers in the "boonies" especially of the coming generation of researchers to be screened and applied where appropriate?

The conference of professionals is one means of transmission. Other devices include the nationwide recruitment by policy agencies of personnel, including interns; the publication in policy journals of summaries of research; and the Washington internship organized around policy problems. Still other means have been tried, including the structuring of meetings and publications by a mix of highly regarded scholars in their field and younger persons whose research is little known, the design of interuniversity study teams, and computerized bibliographical reference materials.

Essentially we must know better than we do, and far more systematically, what works to get important research findings into the policy debate, and it is necessary that we know the processes of formulating transmission means that do not become overloaded. More specifically, we need screening devices that work and that at the same time do not obscure.

--- Selection devices for screening research There is a recognized elitism in the professions, especially in the research communities, that makes for single lines of research and of research findings. Though this phenomenon has been the subject of much discussion over the years as it applied to breakthroughs in medicine--for example, the recorded response of the medical profession to Pasteur's work--there is much less examination of the phenomenon as it occurs in the social sciences. Many factors contribute to such singleness of view. I recall one time in the 1960's when the jockeying for position as Economic Advisor of the Labour Government in the United Kingdom blocked diversity of opinion among scholars qualified for the post. Yet the unusual concensus was on a conceptual problem of human investments and other residual factors in economic growth. Here were 11 scholars, all with a single view on a very complex theory. (25)

The question of who screens whom and for what is difficult. The elitism and the follow-the-leader phenomenon within that elitist structure create much difficulty in a screening program that will encourage needed policy research rather than production of research that is only rewarded because it fits the pattern that is fashionable among scholars.

--- Organization of governments for responsiveness to research--Unless there is a structure for asking for research findings and a "need to know," it is unlikely that research will be sufficiently supported, that the right research questions for policy purposes will be defined, or that research will find its way into policy applications. The structure, organization, and staffing provided by the move toward program analysis and evaluation at all levels of government make it necessary to define problems, search for options, examine optional means for cost and effect, and also to isolate impacts of increments of research use and program services.

At present, however, the Federal evaluations and analysis need beefing up. But Federal encouragement of good analysis and evaluation at state, city, and county levels is even more important. The requirements for evaluation by the states and the local governments in some Federal statutes mark a beginning but far more study is needed to find ways to support those efforts, to buttress them by staffing, and to provide necessary

technical assistance.

It is essential, too, that we know the feedback consequences of requirements for analysis and evaluation that cannot feasibly be met, and that are not now met with the staff resources and support in funds and other assistance that are provided. Requirements that are realistic in size and timing have to be explored for use at each of the levels of government.

--- University support of analytical and evaluation efforts by governments--Few dispute the facts about university inputs into policy. University research has become policy oriented only recently and still in only some places. The second-rate position of applied research within the academic disciplines affects both the role and the quality of policy-directed research. Such research as is done is often viewed by the governments as untimely, unusable in terminology or construct, and expensive compared with alternative means for the responsiveness to public problems gained. In some instances, the financial incentives work to the disadvantage of those who undertake policy studies in terms of pay, promotions, tenure, and optional opportunities, as well as professional respect.

Yet in the health field the universities and in particular the schools of public health have played an important role. They traditionally have been training centers for personnel going into public service and have carried out much public health research, including policy studies. The centers are now being buttressed by other policy centers in universities.

How to structure incentives for the university scholars so that the research is policy oriented and encourages by its usefulness governmental application of policy findings is the primary question for study.

Footnotes

1. Robert Dorfman, ed., Measuring Benefits of Government Investments (Washington, D.C.: Brookings Institution, 1965).

2. S. Mushkin, "Health as an Investment," Journal of Political Economy 70 (October 1962): 129-57.

3. T.C. Schelling, "The Life You Save May Be Your Own," in Problems in Public Expenditure Analysis, edited by Samuel Chase (Washington, D.C.: Brookings Institution, 1968).

4. William Vickrey, Comments on the discussions prior to the publication of Measuring Benefits of Government Investments, edited by R. Dorfman (Washington, D.C.: The Brookings Institution, 1965).

5. H. Klarman, The Economics of Health (New York: Columbia University Press, 1965).

6. A.R. Prest and R. Turvey, "Cost-Benefit Analysis: A Survey," The Economic Journal 74 (December 1965): 683-735.

7. Lester Lave and E.P. Seskin, "Health and Air Pollution," The Swedish Journal of Economics 73 (May 1971): 79-95.

8. Selma Mushkin and Ralph Freidin, Lead Poisoning in Children: The Problem in D.C. and Preventive Steps (Washington, D.C.: Public Services Laboratory, 1971).

9. Selma Mushkin and Ralph Freidin, The Case of Lead Poisoning: Prevention Control in D.C., PPB Note #13 (Washington, D.C.: Public Services Laboratory, 1972).

10. Public Services Laboratory, Transportation as a Component of Health Care in the District of Columbia: An Analysis (Washington, D.C.: Public Services Laboratory, in process).

11. Richard Gill, "Emergency Ambulance Service A & B," in Teaching Cases in Planning Programming Budgeting for State and Local Governments, edited by G. Taylor and R. Gill (Boston: Intercollegiate Case Clearing House, 1969).

12. Personal Communication with Dr. Paul Gertman, Boston University (May, 1973).

13. Personal Communication with Home Care Services of the District of Columbia (October 1973).

14. S. Mushkin, J. Surmeier, J. Kane, and D. Detling, "Federal Funding and Intergovernmental Coordination for Drug Addiction Programs," in Drug Use in America: Problem in Perspective, Appendix, edited by National Commission on Marihuana and Drug Abuse (Washington, D.C.: Government Printing Office, 1973).

15. John Surmeier, "Comprehensive Planning for Community Service: Drug Abuse," in Program Evaluation: An Analysis of Performance vs. Original Plan and Promise (Washington, D.C.: Association for Public Program Analysis, 1972), p. A13.

16. Task Force on Narcotics and Drug Abuse, Task Force Report: Narcotics and Drug Abuse: Annotations and Consultants' Papers (Washington, D.C.: Government Printing Office, 1967).

17. Carol Weiss, "Utilization of Evaluation: Toward Comparative Study" in Readings in Evaluation Research, edited by Francis Caro (New York: Russell Sage Foundation, 1971).

18. Joe Wholey, J. Scanlon, H. Duffy, J. Fuxumodo, and L. Vogt, Federal Evaluation Policy (Washington, D.C.: The Urban Institute, June 1970).

19. S. Mushkin, "Evaluations: Use with Caution," Evaluation 1,2 (1973): 30-35.

20. U.S. President, "The Budget Message of the President" in The Budget of the U.S. Government, Fiscal Year 1974 (Washington, D.C.: Government Printing Office, 1973).

21. Kermit Gordon, ed., Agenda for the Nation (Washington, D.C.: The Brookings Institution, 1968).

22. C. Schultze, E. Fried, A. Rivlin and N. Teeters, Setting National Priorities, The 1972 Budget (Washington, D.C.: The Brookings Institution, 1971).

23. C. Schultze, E. Fried, A. Rivlin and N. Teeters, Setting National Priorities, The 1973 Budget (Washington, D.C.: The Brookings Institution, 1972).

24. E. Fried, A. Rivlin, C. Schultze, and N. Teeters, Setting National Priorities, The 1974 Budget (Washington, D.C.: The Brookings Institution, 1973).

25. John Vaizey, ed., The Residual Factor in Economic Growth (Paris: OECD, 1964).

THE RELATIVE CONTRIBUTIONS OF HEALTH CARE AND SOCIAL FACTORS TO HEALTH
PUBLIC POLICY IMPLICATIONS

Monroe Lerner

The Johns Hopkins University

The work necessary for the preparation of this paper was supported by grants (5 R01 HS 00110 and 5 T01 HS00012) from the National Center for Health Services Research and Development and (5 D04 AH 00076) from the National Institutes of Health, U.S. Department of Health, Education and Welfare, to the Department of Medical Care and Hospitals, and by a grant from the Carnegie Corporation to the Medical Sociology Section of the American Sociological Association under which a sub-committee on Non-Health Services' Determinants of Health Levels has been functioning.

Apparently by common consent, at least among the fraternity of public health and medical care professionals, we seem to be on the verge in this country of a massive re-structuring of our system for the delivery of personal health services. In fact, proposals toward this end--many alternative and some even contradictory to one another--are plentiful in the U.S. Congress currently and, not surprisingly therefore, they provide a basis for the spirited public policy discussion now taking place.

In that discussion, reasoned argument in support of proposals to restructure the system is often stated as follows:

1. The health level of the population of the United States, as measured by its infant mortality rate and/or years of life expectancy (the most common indicators), is substantially below that of other Western, industrialized countries. (This formulation represents the beginning of the classic "social problems" approach to health, i.e., that a "substantial discrepancy" between "ideal" health level and reality is perceived by a "significant collectivity" in the body politic as existing in this country, a discrepancy which is perceived as rectifiable by collective social action. For a more elaborate formulation of this approach, see Lerner, 1971, pp. 296-8).

2. Since the "goal" of any health services' system anywhere is to "produce" health, i.e., to maintain the health of the population at a high level (and/or to improve it), and since the health level of the U.S. population is obviously lower than it should be, the system in this country is obviously deficient. This deficiency in system, in turn, may result from corresponding deficiencies in either (or both) of two factors--the quantity of resources allocated as input to the system, and the structure of the system, i.e., its patterns of financing and/or organization.

3. Since the quantity of resources allocated to the system in the United States is relatively high, i.e., expenditures on health constitute a larger percentage of Gross National Product (GNP) in the U.S. than is the case for most, possibly even all, of the other Western, industrialized countries, therefore the deficiency must, by elimination, result from a corresponding deficiency in the structure of the health services' system. This is especially true because, in several crucial respects, the structure of the system in this country differs from the structure in many of these other countries. The faultily-structured system in this country performs inadequately, it is argued, and its product is, therefore, inadequate.

4. Among the major structural changes currently under consideration, one proposal is to replace present voluntary health insurance arrangements by national compulsory health insurance[♦]. A second is to replace present fee-for-service solo or group medical practices by health maintenance organizations. Still others are to replace present patterns of professional self-regulation within medicine by professional services' review organizations, or to expand the capabilities of public planning agencies. These proposals have been suggested singly and in various combinations.

As merely one recent example, among very many, of this type of argument, consider these remarks by Dr. Jesse Steinfeld, recently Surgeon General of the U.S. Public Health Service:

> "The United States has the best medical research apparatus in the world, the best undergraduate, graduate, and post graduate medical education in the world, and the most modern, best equipped hospitals in the world... Best research... Best education... Best doctors... Best hospitals... What's the problem?
>
> Among developed countries, the United States ranks 12th in life expectancy for women and 27th in life expectancy for men. We rank 15th in infant mortality. But in expenses or annual costs for each citizen for health, we rank 1st.
>
> Obviously, something is wrong. Something is wrong with our health apparatus. And as we examine the health apparatus, we find that there is no system. Its a non-system. Nobody...no group...no governmental body is responsible for research, education, or the quality, availability, and delivery of health care. That is the major problem. Lack of responsibility. Lack of accountability. Lack of a system. Lack of planning.
>
> What we have is high-priced chaos. We have an unplanned, often unresponsive and incredibly

[♦] Fraser (1972) provides some evidence, admittedly crude, that in nine Western, industrialized countries the degree of government participation in health expenditures is unrelated to the level of infant mortality.

wasteful non-system, utilizing far too excessively our limited human, medical, and technical resources. Health care in the United States is a marvel of high cost and inefficiency." (Steinfeld, 1973, pp. 1-2).

However, there are those who refuse to accept the argument that, because the "product" or our health services' system is not adequate compared to others, therefore the "fault" lies in the structure of the system. Usually one or more of four major classes of counter-argument (singly or in combination) are advanced, as follows.

The first rejects the assertion that health in the U.S. is substantially lower than in other Western, industrialized countries. Health, it is asserted, is a multi-dimensional characteristic, much broader than merely the quantity of life, so that indicators of a population's quantity of life--its infant mortality rate and its years of life expectancy--tap merely one dimension of health.[♦] Further, measures of the quantity of life may not be perfectly, or possibly even substantially, correlated with measures of those aspects of the "quality" of life which are related to health, e.g., freedom from physical and/or emotional illness, impairment, or disability, and possession of "social well-being" and even "positive health", however defined.

Along these lines, the assertion is that at some levels, although perhaps in only relatively minor degree in relation to the total, there is some reason to think that the average quantity of life for a population aggregate may actually, under some circumstances and for short periods of time, be negatively correlated with an aggregate's average quality of life, at least as the latter is reflected in freedom from illness, disability, and impairment (Lerner, 1973d). This occurs when case-fatality rates decline, as they have with advances in medical science and technology, especially in recent years, for those illness conditions where survivorship leaves the individual substantially "impaired" rather than completely "cured" of the condition or, in some instances, of its residual effects. Following the line of reasoning developed in this counter-argument, there may be no "fault" with the structure of the health services' system, since its "product" is not necessarily inadequate when compared to others.

The second class of counter-argument, while agreeing that health in the U.S. is lower than it should be, provides an alternative explanation for it. It asserts that lower health here is the consequence of inadequacies in the system which in turn follow from the allocation of an insufficient quantity of resources to it; even though expenditures on health as a percentage of GNP are higher here than in comparable countries elsewhere (or at least as high), nevertheless they should be even higher than they are. Thus we are said to have a shortage of primary physicians or physicians' services, so that greater output from the medical schools is needed or the productivity of physicians should be extended by the introduction on a massive scale of various kinds of physicians' assistants or other kinds of paramedical personnel. Similarly, the quality of medical personnel is said to be inadequate (e.g., because physicians and others providing the great bulk of patient care in this country are unable to keep up with the latest advances in medical technology emanating from this country's great medical centers), so that we need extensive programs of continuing medical education. Finally, we are said to lack an adequate supply of highly specialized medical equipment (renal dialysis units, cobalt machines, etc.) and personnel (to operate this equipment), of emergency medical services (to provide care instantly and on the spot), of outreach services (particularly to serve the under-privileged), and of home health services and skilled nursing facilities. Each of these inadequacies in the system has in common the characteristic that its correction or amelioration requires additional resources, very likely public resources[♦].

Three types of reasoning support this counter-argument by providing an explanation for the requirement of higher expenditures here than elsewhere. One holds that we may be victims of our own success, i.e., that our very success in reducing mortality from the communicable diseases and other illness conditions, especially at the younger ages and mid-life, has resulted in the survivorship to mid-life and the older ages of many with chronic illness, impairment, or disability. For these people, it is argued, the provision of health services can be, and often is, very expensive, especially because of the elaborate and complicated equipment (renal dialysis units, cobalt machines, etc.), surgical procedures (kidney and heart transplants, etc.), and medications required to maintain them. Perhaps we in the United States have invested more heavily than is true elsewhere in this elaborate and complicated equipment and surgical procedures, and in production of expensive medications, but perhaps even heavier investment is required just because of our very success.

The second type of reasoning in support of the need for higher expenditures here argues that a considerable part of our medical care expenditures is accounted for by the massiveness of our clinical and bio-medical research; its products (both technology and trained personnel) are adopted elsewhere at very little cost. But also, the results of this effort often do not appear in the statistics, either because the years of life saved are too few or because the saving is in freedom from pain and discomfort rather than in years of life. This would be true if, here compared to elsewhere, more of the expenditures for medical care are actually used for the treatment of

[♦] Even these two commonly used measures of health, although probably not completely independent of one another, are nevertheless far from perfectly correlated. Each merits independent investigation.

[♦] Additional resources, whether public or private, would obviously have to be diverted from elsewhere, so that other societal needs might be less adequately met than at present.

essentially incurable or irreversible chronic illnesses.

The third type of reasoning supporting this counter-argument holds that it may be the uniqueness of this country's life-styles, when coupled with its wide heterogeneity in cultural patterns and the enormous physical and geographic mobility of the population, which requires that expenditures for health be higher here to provide the same level of health as elsewhere. Persons holding this view argue that our affluence[♦], the relatively high proportion in sedentary occupations, the large amount of motor-vehicle traffic, our high levels of environmental pollution (related to high industrialization and urbanization), our attachment to cigarette smoking, etc., all of these present special hazards and problems, the solution of which requires additional resources. Yet, even though these problems, by default, have become the responsibility of the health services system, additional resources have not been allocated to the system, and its performance, therefore, appears to the observer to be less effective than it may be in actuality.

Persons holding these views are likely to argue also that Americans demand more of their health services' system than is the case elsewhere. For example, Americans insist on a "personal" relationship to their physician, whatever the structure of the system may be, and further they have been, at least in the past, willing to pay the additional costs thereby incurred. Also, perhaps as a consequence of differences in cultural patterns and in family structure, Americans expect their formal institutions to assume a substantial portion of the burden of caring for the chronically ill, whereas elsewhere this may more likely be done within the family; and while expenditures for the sheer "room-and-board" aspect of the care provided by institutions (nursing homes, etc.) appear in the statistics as medical care expenditures, this is not the case when the same care is provided in the family.

[♦]Glazer (1971), citing an earlier statement by Fuchs, argues that although in the past rising levels of living were beneficial to health, in the United States, at least, we may have entered the stage at which this is no longer true. Auster et.al. (1969), studying the relationship of mortality of whites to both medical care and environmental variables by means of a regression analysis across states using 1960 data, found a positive association between high income and high mortality when the effects of medical care and education were controlled for. They speculated that this may reflect unfavorable diets, lack of exercise, psychological tensions, and other factors, and that it may explain the failure of death rates to decline rapidly in recent years. The logic here is that adverse factors associated with the growth of income may be nullifying the presumably beneficial effects of increases in the quantity and quality of medical care. For some further supporting reasoning, see also Lerner (1973d).

The third class of counter-argument, like the second, also accepts the premise of the argument for re-structuring the system, i.e., that the health level of the United States is lower than is appropriate under the circumstances. However, it argues that the remedy lies not in restructuring the system or even in increasing the allocation of resources to it. Rather it argues that, even at present levels of allocation, the performance of the system (i.e., the health level of the population) could be improved materially by changing the "mix" currently making up the health services' system, i.e., the resource allocation pattern presently existing among its various subsystems, in the direction of increasing the share, relative to others, of those providing the largest return on investment. Return, for this purpose, is measured in terms of improvement in the major indicators of system performance in current use in public policy debate, i.e., life expectancy and the infant mortality rate, or whatever other indicators become the vogue. However, the merit of this counter-argument appears to rest largely on the usefulness of the taxonomy of sub-systems and the feasibility of basing public policy on it.

One such taxonomy which appears to merit careful consideration has been advanced by Stewart (1971). He divides the health services' system into four sub-systems[♦] defined by their objectives: treatment, prevention, information, and research. At least two other sub-systems could perhaps be added here, one intended to bring about recovery and rehabilitation, and the other relief from dissatisfaction, pain, and discomfort (this latter perhaps including relief from anxiety about illness). Also, each of the proposed sub-systems could itself in turn be further sub-divided; for example, under treatment a separation might be made between surgical and medical. Thus, one school of thought points to an apparent "over-supply" of surgeons in this country and an apparent "under-supply" of primary-care practitioners (Stevens, 1971). Again under treatment, a separation might be made by place of treatment, e.g., inpatient versus outpatient ambulatory; by type of disorder, e.g., life-and-death situations or conditions including emergencies, chronic progressive conditions, and relatively mild self-limiting diseases (Teeling-Smith, 1973); by age-group of patients (the aged, persons in adulthood and mid-life, children and youth, and infants); or in any other way that seems appropriate.

Finally, the fourth class of counter-argument holds that, even though the "goal" of the health services' system in the United States is to maintain the health of the population at a high level (and/or to improve it), and clearly the system does have an enormous effect on health, nevertheless factors other than the health services' system also exert a substantial influence on the health level of the population (Lerner, 1973a; Benham, 1971; Glazer, 1971). Thus if the health level is too low, it should not be

[♦]Stewart uses the terms "industry" and "primary system" to mean what is designated above as "system" and "sub-system".

attributed, at least not solely, to the structure of the system. Rather, the influence on health of these other factors--factors other than the health services' system--should be investigated and measured, so that their possible manipulation in the interest of improving the health level of the population, either as an alternative to or in addition to changing the structure of the system, should also become a matter for public policy debate.

Although some isolated studies along these lines have been conducted, on the whole this activity--investigation and measurement of the relative influence of the health services' system and of non-health services' factors on health levels--has not previously been carried on in systematic fashion, possibly because of the stridency of the public policy debate centered around changing the structure of the system, but possibly also because no systematic framework has heretofore been available within which discussion of the results of these studies could be located. The present effort is offered as a contribution to what is perceived here as the desirable widening of the focus of that debate. It provides a very preliminary framework for conceptualization of the non-health services' factors which may influence health levels, and it suggests lines along which it may be profitable to pursue further inquiry.

The framework--independent variables

The framework of independent variables suggested here represents a modification and further development of one presented earlier by the present author (Lerner, 1973c) on behalf of a working group established under a grant from the Carnegie Corporation to the Medical Sociology Section of the American Sociological Association to explore the "Non-Health Services' Determinants of Health Levels". That framework divided the factors affecting health levels into two major categories: those "endogenous" and those "external" to the individual.

To quote directly from the earlier report...

> "The endogenous factors consist of five major sub-categories: genetic endowment; biology of the organism; personality factors; cognitive factors; and behavioral patterns. By the last of these sub-categories, behavioral patterns, what is really meant is life styles considered at the level of the individual, i.e., as selected by the individual for whatever motives or under whatever constraints from among the various alternative life-styles available to him. In turn, and for present purposes only, a life-style might be defined operationally to include these components: level of living; type of occupation; food and nutritional habits; degree of social insurance and/or other forms of protection against various types of ecomonic and social insecurity; propensity to use, and the availability of, medical care services; and, finally, personal hygiene habits and patterns.
>
> The factors external to the individual are conceptualized here as including the following sub-categories: the "environment" in which the individual lives and/or in which his "community" is situated; his "community"; the social groups which are significant to each individual and may be hypothesized as directly affecting his health; and finally the system for the delivery of personal health services, but only to the extent here that it introduces iatrogenic factors, presumably "unintended" as detrimental to the health of the individual, but nevertheless in fact having that consequence. Each of these sub-categories in turn merits some further discussion.
>
> The environment includes not only the more obvious physical and biological features impinging on health, but also a very large social component... Much of the advance in health throughout human history has...taken place as a consequence of advance in human ability to modify and control for "social" purposes the physical and/or biological features of the environment, and there really is no reason to suppose that this chapter is closed. Under "social environment" we include culture (conceptualized here as wider in scope than community), and location. Culture, in turn, in our formulation includes values, the state of the arts and the level of technology, and the modes, types, and speed of cultural change.
>
> But individuals live in a community as well as in an environment. Communities, by their very nature, engage to some degree in collective activities, e.g., for the provision of food and other forms of subsistence, maintenance of security and order, integration of moral values, and maintenance of social control. Each of these is crucial to the continued survival of the community and the individuals within it. But communities at a more "advanced" level also engage in public health activities--e.g., disposal of liquid and solid waste; food sanitation activities; water and air pollution control, etc.--and the consequences of each of these activities for health is substantial. Finally, and still under the category of the "community", they provide some sort of a social structure--a stratification system, an occupational structure, etc.--and each of these has ramifications for health...
>
> Within the community--from one point of view, a sub-category of it--individuals are members of, or have reference to, various significant social groups, e.g., their families, other "primary" and various "secondary" groups, formal and informal organizations, residential institutions, etc. In these significant social groups, as defined here, they engage in role relationships and receive support, either positive or negative, through them. The quality and/or quantity of this support is believed by many to be a most significant factor influencing the health of the individual."

For purposes of the present discussion, social factors are defined as including all of the major categories and most of the specific items under "factors external to the individual". But it also includes any endogenous or external factor believed to be capable of being altered, at least in some degree, by collective social and/or political

decision. Thus this definition of factors as social, and yet as affecting health levels, cuts across the earlier categorization. Presumably many, perhaps most, external factors can obviously be modified by collective decision, but presumably also even the endogenous factors which are seemingly "given" (genetic endowment, biology of the organism, etc.), are nevertheless capable of alteration by social and/or political decision (e.g., at least in the sense that, even if they cannot be changed for a single individual at a given point in time, nevertheless their distribution in the population can probably be altered by "eugenic" policies). Thus the concept "social factors", as defined here, is directly relevant to public policy formulation.

The framework--dependent variables

The framework of dependent variables follows an earlier formulation (Lerner, 1973a). Since life is the necessary pre-condition for health, health is perceived in that formulation as a function of both the quantity and quality of life. Quantity of life is measured by life expectancy and (for a population) mortality rates. Although this conflicts with implicit social valuations, each unit of life, regardless of age or stage of development of an individual, is customarily given the same weight in computing life expectancy and mortality rates.

Quality of life, in turn, consists of physical, emotional, and social well-being. States of physical and emotional well-being are related to the presence and/or absence, frequency, and severity of illness, impairment, and disability, with the latter (disability) perceived as the subjective response to objective conditions (illness and/or impairment). Social well-being consists (Lerner, 1973b) of the following sub-components: economic welfare; major-social-role-related coping ability (ability to cope with challenges related to major social roles, lack of dependency, and ability to take advantage of opportunities for personal improvement and development); family health (the health of the family qua family, primarily considered in terms of the social support it provides to the individual to cope with threat, anxiety, illness, etc.); social participation (engagement) in the community (outside one's immediate family) and the quality of personal experience; and perception of moral worth.

The measurement problems here are as yet unresolved. One major problem is the weight to be given in the construction of an aggregative index: 1. to the quantity versus the quality of life; 2. to the various components of the quality of life; 3. under physicial and emotional well-being, to the frequency versus the severity of illness, impairment, and disability; and 4. under social well-being, to its various sub-components as designated above. Clearly, if improvement in the population's health is a desired goal of social policy, some sort of a weighting system, reflecting a consensus of social valuation, should be devised and made explicit, if only for the purpose of making possible the development of a logical framework in terms of which to assess, for public policy purposes, the relative contributions to health of social factors, as here defined, and of the system for the delivery of personal health services.

The framework--the relationship of independent to dependent variables

The independent variables include two broad classes of factors--the health services' system and the "social" factors as here defined. If the objective of social policy is to maximize health, and the quantity of societal resources to be allocated for this purpose is fixed, what is the optimum sub-allocation by factor? To find some answers to this is a major problem for public policy research.

Within the health services' system at current levels of resource allocation, two major types of change are possible. One is to alter its structure, along one or more of the lines indicated earlier. The second is to change its resource allocation pattern, again along lines indicated earlier. It should be noted that neither change is likely to be without cost, even if this is nothing more than the cost of the change itself, i.e., to move the system in the desired direction. Also, many proposed changes combine elements of these two major types. For example, the current HMO proposals suggest structural changes in medical practice (from fee-for-service solo to group practice of medicine) along with changes in the pattern of allocation by sub-system (curative to preventive care, inpatient to ambulatory, etc.)♦.

Insofar as social factors are concerned, again at current levels of resource allocation, there is no system in the same sense as a health services' system exists, and other than the entire economy and social system, the structure of which could be altered to improve health. However, changes in resource allocation patterns among various current social interventions could occur, perhaps paralleling those suggested earlier as applicable to the health services' system. This becomes much more complicated, however, and it seems more likely that additional increments, rather than resource transfers, are likely to be considered.

Where could additional increments of social intervention be expected most efficiently to improve health? One proposal is to invest heavily in reducing air pollution. (For an excellent discussion of benefits, see Lave and Seskin, 1970 and 1972.) Merely some others might be to reduce cigarette smoking and other addictions deleterious to health; reduce accidents of all kinds, but especially motor-vehicle accidents; and encourage weight reduction among the moderately overweight and obese, by improving the population's nutritional intake (Henderson, 1972), and promoting widespread participation in regimes of light to moderate exercise, and in other ways♦♦. Many similar proposals exist.

♦For a most perceptive statement along these lines, see Garfield (1970).

♦♦Each of these policies is, at this point, not widely considered to be the responsibility of the health services' system, but there is considerable support, at least in some quarters, for the sentiment that they ought to be.

Another social intervention is to strengthen the family as a family, e.g., by providing alternatives to institutionalization for the aged and infirm, by expanding the scope and variety of social services available to families, and by providing counselling services to "problem" families or families on the verge of dissolution.

Two points should be noted about these and probably any other major social interventions which might be proposed for the purpose of improving health. One, each is likely to involve direct economic cost of substantial magnitude; and two, no certainty exists, and possibly not even a substantial probability, that each of these interventions is likely to be "successful" in achieving that improvement. Also, those proposed here reflect merely the bias of this author; they may not be those which will be found in practice to provide the greatest health benefit to the community for the least allocation of its resources.

Many, and perhaps all, of these and other social interventions have their counterparts at the level of individual behavior, often involving the lifestyles and the economic choices of individuals. (It is in this sense that the distinction between characterisics endogenous and those external to the individual become less meaningful for purposes of public policy formulation.) For example, the social interventions aimed at reducing the hazards of the physical and/or biological environment usually, if not always, involve motivating and/or educating individuals not to engage in behaviors which expose them to these environmental hazards. Since these behaviors often provide short-term gratifications to those who engage in them--for example, the tension reduction believed to be provided by cigarette smoking, driving at excessive speeds, and "overeating"--and for other reasons, success in health education aimed at altering these behaviors has by and large proven to be elusive; even more elusive has been success in motivating individuals in our society, at least on a large scale, to engage in preventive, "positive" activities beneficial to health.

A complication in using social intervention to improve health is that the social factors to be altered simultaneously serve societal purposes and needs other than health. Any alteration along these lines, even when beneficial for health, may nevertheless as a secondary effect be socially deleterious otherwise. The example that comes most readily to mind here is that the presumably beneficial consequence for the health of the American people which would result from the elimination of cigarette smoking would certainly have a deleterious secondary effect, at least in the short run and unless offset by expensive "rehabilitative" measures, on the economy of tobacco-growing and cigarette-manufacturing North Carolina and elsewhere. Much the same statement can probably be made about most social interventions and/or public health measures, i.e., the health benefit they produce may often, in considerable degree, be offset by their negative secondary effects on the economy or society. Although decision as to the relative weights to be accorded to primary and secondary effects of any intervention should be made by the body politic as part of the political process, this decision can certainly be informed by precise measurement of effects, whether primary or secondary, as part of a research endeavor.

In many instances these forces operate in reverse, i.e., social policies adopted to meet societal needs other than health may also have secondary consequences for health. Here the instance that comes most readily to mind concerns the current energy shortage. This shortage is surely the long-run consequence of many factors but in the short-run most immediately of the decision by the Arab oil-producing nations to curtail oil production; the response in the United States was to reduce both the volume of motor-vehicle use and travel speeds. Very likely, although definitive data are not yet♦ at hand in this matter, a secondary consequence of these steps will probably be a fairly substantial reduction in motor-vehicle deaths, and therefore an improvement in the health of the American population and perhaps even of its international ranking in life expectancy. This was surely not the intent of the oil producers nor even of the U.S. Government. However, other consequences may perhaps also follow from reduced motor vehicle use, deleterious for the economy as employment drops, beneficial for health as air pollution from motor-vehicles declines, deleterious for health as emergency medical services are reduced due to the gasoline shortage, and many others. The ramifications appear endless, and measurement becomes correspondingly complex.

Still another complication in altering social environments to improve health is the distributional consideration, i.e., that the impact of these alterations may be unequally distributed among different segments of a population♦♦. (The same unequal distribution of benefits may, of course, also result from alterations in the structure of the health services' system or from alterations in the pattern of allocation of resources for that system.) This follows from the assumption that the health of any population or population segment is a function of the resources (whether public or private) devoted to the health services' system or other aspects of society serving it; thus, since resources are finite, the distribution of health among populations approximates a zero-sum game. That is, at any given level of resource allocation, if the level of health for one population is to be relatively "high", the corresponding level for another will have to be relatively "low". To the degree that social policy determines resource allocation, and to the degree that resource allocation influences health, the relative health of populations represents the results of policy decisions related to resource allocation.

Perhaps the best way of stating the public policy question here is not to ask about the relative

---

♦ December, 1973.

♦♦ For a perceptive discussion along these lines, see Rivlin (1971, pp. 56-60).

influence of social and health services' system factors on health, but rather to ask about the relative influence on health of marginal increments or decrements of resources allocated to either of these general factors or to their specific components. The question then becomes: Given the availability of an additional dollar (or, conversely, if we have to take one away), where should that dollar be invested to produce the maximum increase in health? And here the difficulty is that, while at least theoretically even the very diverse factors comprising the independent variable as here conceptualized can ultimately be expressed in terms of dollar costs to the community, this is not (perhaps not yet) true of the dependent variable, health. No one has constructed a generally acceptable index to represent health in which comparable units of the quantity and quality of life could be aggregated into a single number, ultimately the number of dollars♦.

♦Robertson (1971) uses work-loss as an indicator of health, but even this approach does not quantify the non-income benefits of health.

BIBLIOGRAPHY

Auster, Richard, et.al.
1969 "The Production of Health: An Exploratory Study". Journal of Human Resources 4,4: 411-436, Fall.

Benham, Lee
1971 The Effect of Medical Services on Health. Center for Health Administration Studies, University of Chicago. Processed.

Fraser, R.D.
1972 "Health and General Systems of Financing Health Care". Medical Care 10,4:345-356, July-August.

Garfield, Sidney R.
1970 "The Delivery of Medical Care". Scientific American 222,4:15-23, April.

Glazer, Nathan
1971 "Paradoxes of Health Care". The Public Interest 22:62-77, Winter.

Henderson, L.M.
1972 "Nutritional Problems Growing Out of New Patterns of Food Consumption". American Journal of Public Health 62,9:1194-98, September.

Lave, Lester B., and Eugene P. Seskin
1970 "Air Pollution and Human Health". Science 169,3947:723-733, August 21.

1972 "Air Pollution, Climate, and Home Heating: Their Effects on U.S. Mortality Rates". American Journal of Public Health 62,7: 909-916, July.

Lerner, Monroe
1971 "Health as a Social Problem". In E.O. Smigel, Editor, Handbook on the Study of Social Problems. Chicago: Rand-McNally, pp. 291-330.

1973a "Conceptualization of Health and Social Well-Being". Health Services Research 8,1:6-12, Spring.

1973b "Indicators of Social Well-Being". An unpublished paper for presentation at a conference sponsored by The Rand Corporation. July 21. Processed.

1973c "The Non-Health Services' Determinants of Health Levels: Conceptualization and Public Policy Implications". Paper delivered at Annual Meeting, American Sociological Association. Processed.

1973d "Modernization and Health: A Model of the Health Transition". Paper delivered at Annual Meeting, American Public Health Association. Processed.

Rivlin, Alice M.
1971 Systematic Thinking for Social Action. Washington, D. C.: The Brookings Institution.

Robertson, Robert L.
1971 "Economic Effects of Personal Health Services: Work Loss in a Public School Teacher Population". American Journal of Public Health 61,1:30-45, January.

Steinfeld, Jesse
1973 "Position of Dr. Jesse Steinfeld", in: Health Politics, A Quarterly Bulletin. Committee on Health Politics, New York University. 4,1:1-4, October.

Stevens, Rosemary
1971 "Trends in Medical Specialization in the United States". Inquiry 8,1:9-19, March.

Stewart, Charles T., Jr.
1971 "Allocation of Resources to Health". Journal of Human Resources 6,1:103-122, Winter.

Teeling-Smith, George
1973 "More Money Into the Medical Sector: Is This The Answer?" International Journal of Health Services 3,3:493-500.

Discussant: Berton H. Kaplan, Ph.D.
Department of Epidemiology
University of North Carolina School of Public Health
Chapel Hill, North Carolina

Summary:

All the papers in this section deal with a common question: What is a health-effective social system? The panel is a Durkheimian celebration: How do social bonds affect health status?

Professor Monroe Lerner's (Ph.D.) paper poses a number of challenges: (1) The need to project futurist implications of health delivery systems on a national, class, and/or ethnic basis; (2) The need to project futurist implications of health policy; (3) The need for looking at new ways of strengthening the social system - to learn to modify deleterious social situations as part of prevention.

Professor Sidney Cobb's (M.D.) paper challenges us to consider how overload situations can be altered as a preventive strategy. Overload is one way of looking at a number of disease effects through a basic "stressor" of modern life - overload, overstimulation, over-achievement heroics. Overload is also perhaps the Protestent Work Ethic in the act of demonstrating worth.

Professor Selma Muskins (Ph.D.) paper illustrates a basic contribution of the poor - a re-evaluation of health policy and health care. Her research is also a challenging illustration of the evolution of a "right" health care. Her work is policy in the concrete, not as abstract "models". Lead poisoning is real. Policy leads to feedback, to evaluation of effectiveness.

It occurs to me that this panel can be summarized around the need to develop a Leontieff type input-output matrix of the health services system, with an elementary set of "cause" or risk inputs to the basic set of care-taking-preventive parameters of the process of level of health states.

STUDENTS' ATTITUDES TOWARD MATHEMATICS AND STATISTICS IN THE SOCIAL SCIENCES

Ibtihaj S. Arafat, City College of City University of New York

Introduction and Review of Literature

As an applied science, statistics has necessarily two aspects, one local and the other universal. The theory of statistics supplies the foundations of inductive inference and, in doing so, it uses the calculus of probability and other Mathematical apparatus. Statistics thus includes in its scope the most abstract mathematical work: principles and methods relating to the design of experiments and surveys, the drawing of valid inferences from observational data, and a very large volume of applications in many different fields of specialized activities. On the other hand, statistics, in its present form, is almost a name for a group of subjects which have often developed independently and some of which still remain, in many ways, separate disciplines. Statistics, in the form of observational data, constitute the subject matter of natural and social sciences and also supply the basis for decisions in administration and in social and economic affairs (Mahalanobis, 1957:14).

For the purpose of this study, attitude is defined as the respondents' enjoyment, interest, and to some extent their level of anxiety towards the subject matter. In other words, an attitude is a tendency to act in a certain way. Mathematics is defined as the science that deals with the relationships and symbolism of numbers and magnitudes. It includes quantitative problems, including all operational functions dealing with numbers and computer science. And the social sciences are defined as that branch of science that deals with the institutions and functions of human societies and with the interpersonal relationships of individuals as members of a society. This includes psychology, political science, sociology, and economics.

This study examines some of the factors affecting social science students' attitudes toward mathematics and statistics, since statistics uses a great deal of mathematics. But statistics are also applied to various other sciences, to the point that the field has developed into many separate disciplines. Statistical methods have been found useful for solving the many problems that underly the task of providing facts as a basis for social theory.

Iversen (1972:145) expressed the opinion that the relationship between statistics and the social sciences is not without problems. He stated that, "The statistical profession, or perhaps more precisely, the academic branch of the profession, has too often placed itself within a limited framework and become a branch of mathematics. From that niche it has frequently imposed mathematical theorems of restricted values on the social sciences; the theoretical results are seen by statisticians as interesting in themselves and were perhaps never intended to be applied. But, where they have been, the statistical profession has taken little responsibility for the maintenance and none for the uses of the methods."

The key word here is lack of "responsibility." The method of teaching of statistics must be revised since it plays a significant role in the development of the attitude towards the discipline. Lack of knowledge about the subject of statistics, its functions, and its values may also have contributed to the development of a negative attitude toward the field. In other words, previous high school preparation in math, the number of courses taken, and the effects mathematics teachers have on students affects their attitudes towards statistics in later years (Brown, 1933:43, College Entrance Exam Board, 1970). Aiken (1970) states that, "The conception of a typical mathematics teacher as grim, brutal, dull, uncaring and ineffective (which is held by a sizable percentage of students with negative attitudes) may be a matter of "sour grapes." Nevertheless the degree of teacher understanding, effectiveness, and appreciation of mathematics, and particularly those of the most recent teachers, are significantly related to student attitude. Phillips (1970) has noted that improving teacher attitudes towards mathematics can result in a more positive attitude on the part of students. Meinkoth (1971:69–72), concluded in her study that the individual instructors do not significantly influence the achievement of their students but do influence dramatically the attitudes of their students, not only toward the instructor but toward the subject matter of study. Aiken (1970) observed that a student's attitude toward mathematics would be important in determining whether he elects to take courses in mathematics. And in later study he (Aiken, 1972) stated that there is a positive relationship between students' attitudes toward mathematics teachers and taking math courses in later years. Bending and Hughes (1954) expressed the opinion that increased exposure to mathematics courses in high school and college gives the student a familiarity with and a liking for intellectually handling the quantitative concepts so important a part of statistics.

Banks (1964) observed that an unhealthy attitude toward arithmetic may result from a number of causes. By far the most significant contributing factor, however, is the attitude of the teacher. The teacher who feels insecure, who dreads and dislikes the subject, for whom arithmetic is largely rote manipulation, devoid of understanding, cannot avoid transmitting his feelings to the student. On the other hand, the teacher who has confidence, understanding, interest and enthusiasm for arithmetic has gone a long way toward insuring success. Leak (1970) states that there is no relationship between sex or major area of study and interest in the subject matter on the part of the student. On the other hand, Edger (1965) hypothesized that student personality variables–need for achievement, economic and theoretical orientations, and anxiety in testing situations–were factors contributing to student performance. In males, anxiety in testing situations and anxiety over need for achievement have no relationship to their performance. The economic orientation of a male student as a breadwinner generally has an effect on his performance. For females, the more theoretical her orientation, the better her performance. Her anxiety about achievement, her anxiety in testing situations, and her type of economic orientation, are all variables contributing to her performance. Males are not affected by these variables. This study suggested that there would be differences between male and female attitudes.

The review of literature on the topic indicates that many studies were conducted by researchers who concerned themselves with attitudes towards mathematics and statistics as a reflection of students' early experience with the subject, their teachers, and their individual differences in personality. And as has been shown above, researchers have varied in their conclusions. This study is intended to cover aspects which have not been investigated by other researchers. It is the first study that is concerned with the attitudes of social science students toward statistics and mathematics.

Methods and Procedures

This study was designed to survey social science students' present attitudes toward mathematics and statistics, and to locate factors in the students' past that might account for these attitudes. Data were gathered by means of a self-administered questionnaire. The questionnaire was distributed to 1290 social science students at an urban university campus in the northeastern part of the United States. 781 (60.5%) of the questionnaires were used and 509 (39.5%) were discarded because they were not complete or lost. The questionnaires were distributed with the cooperation of the sociology department (Chairman and staff) on the above-mentioned campus. About 65% of the respondents were social science majors. The data were collected during the spring session of 1973. The questionnaire consisted of 21 questions, many of which had five or more parts. Questions one through seven covered general demographic characteristics of the respondents, namely: age, sex, ethnic origin, religion, father's occupation, mother's occupation, and family income. Questions eight and nine asked the students about their classification in school, and their major field of study. Question 10 asked the respondents to state how important they believe knowledge of mathematics and statistics is to their present major. The following three questions (11–13) included questions about the kind of high school the students graduated from, the name of the school, the date of graduation, how many courses they took in high school in arithmetic, algebra, geometry, calculus, trigonometry, statistics, and the total number of mathematics courses. Question 14 asked the students about their major area of interest in high school: natural sciences, social sciences, humanities, mathematics, and statistics. The following question (15) asked the students to check and specify the areas required from the above-mentioned areas. Then they were asked (question 16) to rate the quality of teaching in each of the above-mentioned areas on an eight point scale from very poor to very good. Question 17 asked how much the students liked their high school teachers in the same five areas mentioned above. An eight point scale was used to place responses. The 18th question asked students if they had a choice between taking a course with a lot of reading but no statistical or mathematical manipulations, and a course with statistical and mathematical manipulations, which one would they choose. Question 19 consisted of 28 parts, where the students were asked to use a scale of eight points, starting with point one as dissimilar, and point eight as similar, to rate the following college majors: humanities, statistics, biology, physics, psychology, sociology, and economics (28 combinations of majors) according to how similar or dissimilar they believe these fields are. Question 20 asked the respondents to rate each of the following college majors: natural sciences, social sciences, humanities, mathematics, and physical sciences, on a six point scale with very low as one and very high as six, according to how much they believe the particular major contributes to each of eight specified areas, namely: social good, help you lead a better life, usefulness in job, help in understanding public issues, help you make more money, help in developing personality, help in developing mental ability, and help in developing personal rationality and logic. The last question (21) asked respondents to state in their own words the main factors, which they believe affected their attitudes toward mathematics and statistics.

The Chi-Square test, attitudinal scales and percentages were used in the analysis. The method of statistical analysis used served the purpose of revealing patterns of relationships between the variables which are not immediately evident upon examination of the contingency tables.

Analysis and Discussion

The sample (781 respondents) consisted of 51.2% males and 48.8% females. Of the males, 63.2% were white, 15.5% were black, 10.9% were

of other social origins. As for the females, 48.9% were white, 27.7% were black, 12.5% were Puerto Ricans, 5.7% were Orientals, and 5.2% were of other origins such as Indians and Middle Easterners. Of the sample total (781 respondents), 16.4% were Protestants, 31.4% were Catholics, and 6.7% were of other religions such as Islam and Buddhism and Hinduism. No significant difference was found by sex and the occupation of the respondent's father. However, the distribution showed that 11.8% of the respondents' fathers were professionals, 21.3% were businessmen, 5.4% were clerical, 23.4% were skilled laborers, 17.8% were semi-skilled laborers, 4.6% were retired, and 8.7% with no occupation or unemployed at the time of the study. As far as the respondent's mother's occupation is concerned, 10.7% of the mothers were professionals, 8.1% were businesswomen, 14.6% were clerical, 7.5% were skilled laborers, 3.3% were house aides, 44.1% were housewives, 7.5% were semi-skilled laborers, 2.0% were retired, and 2.1% were out of jobs at the time of the research. As for the family income of the respondents, 13.0% came from families with annual incomes less than $4,999.99, 17.9% came from families with annual incomes of $5,000–7,999.99, 13.5% came from families with annual incomes of $8,000.00–9,999.99, 19.5% came from families with annual incomes of $10,000.00–11,999.99, 13.6% came from families with annual incomes of $12,000.00–14,999.99, and 22.5% came from families with annual incomes greater than $15,000.00. No significant difference was found to exist by sex and the annual family income of the respondents. The above distribution indicates that our respondents were predominantly middle class.

The educational level of the respondents showed no difference by sex. However, the distribution of data on the different levels showed that, of the sample 18.1% were freshmen, 20.6% were sophomores, 31.5% were juniors, 26.2% were seniors, and 3.7% were graduate students. The distribution by major field of study showed a significant difference between male and female respondents. Of the 5.5% whose major was in the natural sciences, 61% were males and 39% were females. Of the 6.9% who majored in humanities, 53.8% were females. Of the 3.1% who majored in mathematics, 56.5% were males. In physics, out of the 3.6% physics majors in the sample, 66.7% were males. Of the 4.9% of the architecture students, only 18.9% were females, while none of the 3.2% engineering students was a female. This was true also of meteorology, philosophy, and music. Twice as many males as females were majoring in physical education, computer science, pre-medicine, business administration, pre-law, and economics. On the other hand, twice as many girls were majoring in history, education, psychology, art, language, and speech. 94.3% out of the 4.7% of respondents whose major was nursing were females. However, the social science majors, who made up 48.9% of the sample, and the English majors, who made up 2.1% of the sample, were female and male in almost equal numbers. The above distribution indicates a highly significant relationship ($P < .001$) between sex and major field of study.

As for the importance of the knowledge of mathematics and statistics to the major field of study of the respondents at present as perceived by the respondent, a significant difference ($P < .02$) was found to exist between male and female respondents. While 63.2% of the males and 53.2% of the females thought it is important, the distribution on the scale of importance showed that almost twice as many boys as girls believed that it is very important for their field compared to the females, who thought that it is somewhat important. However, almost an equal number (23.4%) of both, the males and the females, believed that it is not important at all for their major field of study.

Most of the sample (94.7%) attended a high school in the New York metropolitan area. However, the sample respondents were the graduates of sixteen high schools in the area.

In answering the question concerning the choice between taking (in college) a course with a lot of reading but no statistical manipulations and a course with statistical and mathematical manipulations, a highly significant difference ($P < .001$) was found to exist between the sexes' choice. While 46.3% of the males would have chosen a course with mathematical and statistical manipulations, only 30.1% of the females would have chosen such a course.

The sample represented the ethnic structure of students in the country at large and the New York metropolitan area in general. 56.2% of the respondents were white, 21.5% black, 11.7% were Puerto Ricans, 10.6% were Orientals and of other nationalities such as Indians, Turks, Middle Easterners, etc. No significant difference was found between ethnic origin and the choice of a course with or with no mathematical and statistical manipulations. However, the Orientals have almost equal respondents in both categories.

As far as the relationship between religious affiliation and income of the families of the respondents is concerned, a significant difference ($P < .001$) was found between the different groups. Most of the Jewish respondents come from upper income level families. Most of those who did not identify with any religion, and those who identified with Oriental religions come from low income families. Protestant and Catholic respondents come from middle range families.

The father's occupation was found to have some effect on the student's attitude towards mathematics and statistics. 51.3% of the respondents whose fathers are professionals felt that mathematics and statistics were very important for them, and 15.6% of the same group felt they were somewhat important. On the other hand, 51.3% of the respondents whose fathers' occupation was clerical felt statistics and mathematics were not important at all and 7.7% of the same group felt it is not important. This implies that the father's educational level affects the children's attitude in general. Also, the mother's occupation had some effect on the respondents' attitudes. The highest percentage of respondents who stated that they would prefer to take a course with mathematical and statistical manipulations rather than a reading course with no statistical and mathematical manipulations have mothers who are working as aides in professional homes or in semi-skilled jobs. These mothers identify with their employers and try to bring up their children in the same manner.

A significant finding is that the graduate respondents are more likely to take a course with statistical manipulations. This reflects their need for the knowledge of statistics. Also a very highly significant relationship was found to exist ($P < .001$) between the major field of study of the respondent and his choice of a course with mathematical and statistical manipulations. The respondents whose major field of study is in the natural sciences, mathematics, physics, architecture, computer science, engineering, pre-med, business, etc., stated their preference for a course with mathematical and statistical manipulations, while most of those majoring in the social sciences, humanities, history, education, nursing, etc., stated their preference for a course with much reading and no mathematical and statistical manipulations. Another significant relationship ($P < .001$) was found to exist between the sex of the respondent and the choice of a course with mathematical and statistical manipulations. This is a result of the way the parents bring up their children and teach them how to play their sex role. Another group that stated their preference for a course with mathematical and statistical manipulations were the Orientals and the others, which included Middle Easterners and Indians. This is probably a result of a language difficulty on the part of these students who find it difficult to take courses with many and varied readings. They try to avoid this problem by taking mathematics and statistics, which do not require a good knowledge of the English language.

To determine the relationship between the number of mathematics and statistics courses in high school and the major field of study in college, the respondents were asked to state how many courses of mathematics and statistics they had taken in high school. Table 1 illustrates the relationship between the number of courses taken in all mathematics courses in high school and the major field of study in college. Of the total sample, 23.6% took one course in arithmetic during high school, 7.04% took two courses, 2.7% took three courses, 2.4% took four courses, and only 1.4% took five courses. Of those students who took courses in algebra in high school, 55.8% of the sample took one course, 29.4% took two courses, 4.0% took three courses, 2.2% took four courses, and 0.6% took five courses. As for geometry, 67.3% took one course, 19.8% took two courses, and 1.5% took three courses. In calculus, 24.2% took one course, 3.0% took two courses, and only 0.5% took three courses. 58.3% of the sample took one course in trigonometry, and 8.7% took two courses, but none took more than two. When it comes to statistics in high school, only 6.0% took one course and 0.6% took two courses only. However, 8.83% of the sample took other mathematics courses in high school.

The above figures indicate that more high school students take geometry, trigonometry and algebra than arithmetic and calculus, and only very few take statistics. The distribution by major field of study in college (Table 1) shows that the respondents took at least one course of mathematics in high school regardless of their major field of interest in high school. However, Table 2 shows a strong relationship between the number of mathematics courses taken in high school and the attitude toward the subject in later years. The higher the number of courses taken, the greater the percentage of those with positive attitude and vice versa.

Table 3 illustrates the relationship between the rate of interest in a certain field in high school and the major field of study in college. A scale of 8 points was constructed for this purpose, with one indicating no interest at all and 8 indicating most interest. The respondents who were majoring in the social sciences in college stated their interest in the social sciences in high school as follows: 21.6% (85 respondents) stated that they were not interested in the social sciences in high school compared to 78.4% who stated different degrees of interest but mainly most interested. When these same students were asked about how much interest they had toward mathematics and statistics in high school, the result was reversed. 55.7% stated no interest in mathematics in high school and 86.7% stated no interest in statistics in high school. In both cases females outnumbered the males. In comparison, when the mathematics majors in college were asked to state how much interest in the social sciences they had in high school, most of them stated either no interest or very little interest. But when they were asked about their interest in mathematics and statistics only one respondent stated no interest in mathematics and five respondents stated no interest in statistics in high school. The interest in the subject matter in high school affects the attitude of the respondent in college toward the subject matter.

The next question investigated the quality of training in the different fields in high school as perceived by the respondents in the different fields of major in college. A scale of 8 points was constructed for the purpose with 1 as very poor and 8 as very good. Of the students majoring in the social sciences in college, 32.2% rated the social sciences teaching in high

school as poor (Table 4) and 67.8% rated it as good to very good. When the same students were asked to rate mathematics and statistics training in high school, 31.6% rated mathematics training as poor, and 77.0% rated statistics training in high school as very poor. When the students majoring in mathematics in college were asked the same question about their social science training in high school, 40.0% of them rated it as very poor, while only 12.9% rated mathematics training in high school as poor, 42.1% rated statistics training in high school as very poor. This shows that the quality of training in a certain field of study in high school had some effect on the students' attitudes toward that field later in their choice of college major.

The next question dealt with the degree of liking of the teachers of the different fields of study in high school as perceived by the respondents, by major field of study in college. Table 5 illustrates this relationship. Of the students majoring in the social sciences in college, 33.8% disliked their social science teachers, 39.4% disliked their mathematics teachers in high school and 68.2% disliked their statistics teachers in high school. In comparison, of the students who took mathematics as a major in college, 44.4% disliked their social science teachers in high school, and only 18.2% disliked their mathematics and statistics teachers in high school. These findings reflect the relationship between how much positive feelings of students toward their high school teachers affects their choice of that field as a major in college.

A scale of 6 points, ranging from very low to very high was constructed to see how much the respondents believed the social sciences and mathematics as college majors contributed to the well being of the individual. Eight areas were defined as components of well being: social good, help in leading a better life, useful in job, help in understanding public issues, help in making money, help in developing the personality of the student, help in developing mental ability, and help in developing personal rationality and logic. The answers of the social science students vs. the answers of the mathematics students were tabulated separately. Table 6 shows the differences between the two groups, and between the males and females within the same group. As for the first point, that is whether the social sciences contribute to the social good, a higher rating was given by the social science majors to this point than the mathematics majors by both sexes. When the same question was asked about mathematics, a higher rating was given by the mathematics majors than the social science majors. The same belief applies to the second point, i.e., if either one of the two fields helps the individual lead a better life. The two groups of students, namely, those majoring in the social sciences and those majoring in statistics, believe that knowledge of mathematics helps in the job the individual takes after he leaves college. On the other hand, the mathematics majors did not think that social sciences are as useful on the job, while the social science majors think that it is. As far as the rating of the point that either one of the two majors help in understanding public issues, again the two groups agreed that mathematics does not help in understanding public issues; while the social science majors believe that the social sciences help in understanding public issues, the mathematics majors do not agree. Both groups have the same low rating concerning the point that majoring in the social sciences helps the individual make more money. However, only the mathematics majors gave a high rating for the effect of majoring in mathematics on making more money in the future, while the social science majors did not believe this is true. Both groups gave high ratings for the effects of the social sciences on the development of the personality. The results were reversed when the effects of both disciplines on the development of mental ability on the individual is concerned. Both groups gave high ratings for the effect of mathematics on the development of mental ability. But social science and mathematics majors disagreed on the effect of the social sciences on the development of mental ability in the individual. While the social science students gave it a high rating, the mathematics students gave it a low rating. The same result was obtained from rating the effects of both disciplines on the development of personal rationality and logic.

When the respondents were asked about the main factors they believed to have affected their attitude toward mathematics and statistics, a number of those who are social science majors gave reasons related to the inadequacy of high school training, and the way it was taught to them by their mathematics teachers. Another reason given by some respondents was "My math teacher was cold. I never liked to be around her." Another respondent stated that "Math is a very dry topic, no talking or discussion by the teacher. We used to be happy when she was absent." Most of the negative factors given by the respondents who held negative attitudes toward mathematics and statistics dealt with their conceptions of their mathematics teacher in high school as dull, grim, does not relate to me, and not interested in the subject in general. On the other hand, the favorable factors given by some respondents were many. One stated, "It is a hard field to do well in, but you get a sense of accomplishment when you understand a difficult problem. My math teacher made me feel that way." Another respondent wrote, "My teachers in math were good. They went out of their way to help me understand the subject," . . . etc. These answers, whether they were negative or positive, reflected theeffect of the high school mathematics teacher on the attitude a student develops toward the field in college.

Implications and Conclusions

The results of the analysis show that 43% of the sample had a negative attitude towards mathematics and statistics, 25% were indifferent, and only 32% had a positive attitude (Table 2). This is approximately a 2:1 ratio against mathematics and statistics. Lack of knowledge about the subject of statistics and mathematics, its functions, and its values may have contributed in developing a negative attitude. However, the findings obtained from this study showed a strong relationship between a number of factors and the attitude of students towards mathematics and statistics. These relationships could be summarized as follows:

1. Previous high school preparation in math and statistics had a strong relationship to the attitudes the student had in college towards the subject.
2. The number of courses taken in high school, and the quality of training had a strong effect on the student in later years. Also the interest in the subject in high school affects interest in later years.
3. The relationship of the student and teachers' interest in the subject matter, their personality and exposure, and how much liking they elicited from students had a strong effect on students' attitudes in later years.
4. A strong relationship was found between attitudes toward mathematics and statistics and gender. Twice as many males as females major in computer science, pre-medicine, business administration, pre-law, economics, engineering, architecture, mathematics, physical education, and physics. This is a result of the early orientation of the males and females generally. Also, a higher number of males than females indicated their choice of taking a course with mathematical and statistical manipulations rather than taking a course with many and varied readings. The author believes that the economic orientation of the male generally had an effect on the orientation toward mathematics.
5. The level of the students in college affects their attitudes also. The graduate respondents are more likely to take statistics and mathematics because of a felt need for it in their advanced studies.
6. Father's and mother's occupations have some effect on the student's attitudes towards mathematics and statistics.
7. The last question indicated that the attitudes and beliefs of the individual are affected by his perceptions about the value of knowledge in a certain area has on his well being, and how much appreciation he develops towards that specific field.
8. Students of certain ethnic groups were found to have a better attitude toward mathematics and statistics. Perhaps this is because they do not need a good knowledge of the English language to study mathematics and statistics. In many instances, the language barrier stops them from performing well in heavy reading courses.

In summary, it was found that the individual's attitude toward mathematics and statistics is not influenced by age, religion, or income. The factors that play a part in determining one's attitude toward the subject are major field of interest, sex, ethnic origin, the attitude one holds towards high school mathematics teachers, the number of high school mathematical courses taken, and the importance of mathematics to one's major. Most of these findings agree with the findings of references cited in the review of literature with some additions and modifications.

Since this study is the first one in this area, the author suggests further research should be undertaken to meet the problems mentioned above. A better screening and preparation of mathematics and statistics teachers should be taken into consideration and investigated so that the subject could be made more appealing to students. A different approach may yield more favorable results in student attitudes and aptitudes.

Bibliography

Aiken, Lewis R., "Research on Attitudes Toward Mathematics," *Arithmetic Teacher,* Vol. 19, March 1972.

——, "Biodata Correlates of Math Attitudes," *School Science and Mathematics,* Vol. LXII, No. 5, 1970.

Banks, John J., *Learning and Teaching Arithmetic.* (2nd edition), Allyn and Bueon: Boston, Mass., 1964, pp. 116–117.

Bendig, A.W., and Hughes, J.B. III, "Student Attitude and Achievement in a Course in Introductory Statistics," *Journal of Educational Psychology,* Vol. 45, 1954.

Brown, Ralph. *Mathematical Difficulties of Students of Educational Statistics.* Teachers College, Columbia University: New York, 1933, pp. 48–49.

Edger, Leo E., "A Multivariate Analysis of Factors Contributing to Students' Performance in the Introductory Course of the Business Administration Curriculum," (Unpublished Masters Thesis), University of Washington, 1965, pp. 1–268.

Editorial Staff of the College Entrance Exam Board, "Survey of Secondary School Mathematics," *School and Society Magazine,* (the survey was conducted by the College Entrance Exam Boerd), April, 1970.

Iversen, G.R. "Social Science and Statistics," *World Politics Magazine.* 1972. pp. 145–54.

Leake, Charles R., "Interest Changes in Mathematics of Selected College Students in New York State," *Dissertation Abstracts International,* No. 30–7–8, January-February, 1970.

Mahalanobis, P.C., *The University Teaching of Social Sciences Statistics,* (edited for The International Statistical Institute in the Hague and sponsored by The UNESCO, 1957, p. 14.

Meinkoth, Marion R., "Teachers of Economic Principles Effect on Student Achievement and Attitudes," *Journal of Experimental Education.* Vol. 40, 1971, pp. 69–72.

Phillips, Robert B., "Teacher Attitude as Related to Student Attitude and Achievement in Elementary Mathematics," *Microfilm,* University of Virginia, 1970.

TABLE 1: THE NUMBER OF MATHEMATICS AND STATISTICS COURSES TAKEN IN HIGH SCHOOL BY MAJOR FIELD OF STUDY IN COLLEGE

No. of Courses Taken in High School in Mathematics		Major Field of Study in College of Students Who Took Mathematics and Statistics in High School														Total
		Natural Sciences		Social Sciences		Humanities		Mathematics		Physical Science		English		Other Areas		
		M	F	M	F	M	F	M	F	M	F	M	F	M	F	
Arithmetic	1 Course	14	14	54	55	5	6	1	2	11	4	1	1	11	5	184
	2 "	3	2	17	13	2	3	4	3	5	0	0	1	1	1	55
	3 "	0	3	8	4	1	1	0	0	1	0	0	1	2	0	21
	4 "	3	2	4	6	1	0	1	0	2	0	0	0	0	0	19
	5 or more	2	1	2	4	0	1	0	0	0	0	0	0	1	0	11
Algebra	1 Course	34	34	119	131	21	20	8	4	25	3	3	4	19	11	436
	2 "	14	14	62	77	6	9	9	5	15	4	4	4	2	5	230
	3 "	3	4	11	6	1	0	1	1	1	0	1	0	2	0	31
	4 "	2	0	6	4	1	1	0	1	1	0	0	0	1	1	17
	5 or more	1	0	1	2	0	1	0	0	0	0	0	0	0	0	5
Geometry	1 Course	44	39	133	159	25	25	13	7	31	3	8	7	18	14	526
	2 "	9	8	51	49	3	5	5	5	11	2	0	0	4	3	155
	3 "	1	2	1	3	1	1	0	0	1	0	0	1	1	0	12
	4 "															
	5 or more															
Calculus	1 Course	19	13	55	41	10	6	8	5	17	1	2	2	5	5	189
	2 "	2	0	8	5	0	1	2	2	3	0	0	1	0	0	24
	3 "	0	0	1	0	0	0	0	0	3	0	0	0	0	0	4
	4 "															
	5 or more															
Trigonometry	1 Course	41	29	124	124	25	20	13	10	29	3	7	5	15	10	455
	2 "	9	3	17	22	0	2	3	1	6	1	0	1	2	1	68
	3 "															
	4 "															
	5 or more															
Statistics	1 Course	7	4	9	7	2	2	2	3	8	0	1	0	0	2	47
	2 "	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2
	3 "	1	0	2	0	0	0	0	0	0	0	0	0	0	0	3
	4 "															
	5 or more															
Other Math Courses	1 Course	3	2	14	12	2	1	2	1	1	1	2	0	2	0	43
	2 "	0	1	2	4	0	0	1	1	0	0	0	0	0	0	9
	3 "	1	0	1	0	0	0	0	0	0	0	0	0	1	0	3
	4 "	1	0	1	3	0	0	0	0	0	0	0	0	0	0	5
	5 or more	1	1	3	1	1	0	0	0	1	0	0	0	1	0	9

* The Grand Total = 781 Respondents = 100.00%
Males Total = 400 " = 51.20%
Females Total = 381 " = 48.80%
M refers to Male
F refers to Female

TABLE 2

THE RELATIONSHIP BETWEEN THE NUMBER OF HIGH SCHOOL MATHEMATICS COURSES AND THE RESPONDENTS' ATTITUDES TOWARDS MATHEMATICS AND STATISTICS

Number of Courses Taken	Positive Attitude	Negative Attitude	Indifferent	Percent in Total
1 course	13%	47%	40%	3.07%
2 courses	14%	64%	22%	8.60%
3 courses	27%	50%	23%	22.75%
4 courses	34%	37%	29%	30.12%
5 courses	34%	46%	20%	12.10%
6 or more	44%	35%	20%	23.36%
Percent in Total	32%	43%	25%	100.00%
				= 781

TABLE 3: MAJOR OF INTEREST IN HIGH SCHOOL BY MAJOR FIELD OF STUDY IN COLLEGE

Major Field of Interest	Rating of The Major Field	MAJOR FIELD OF STUDY IN COLLEGE: Natural Sciences M	Natural Sciences F	Social Sciences M	Social Sciences F	Humanities M	Humanities F	Mathematics M	Mathematics F	Physical Sciences M	Physical Sciences F	English M	English F	Other M	Other F	Total
Natural Sciences	No Interest 1	3	4	17	27	3	4	1	0	2	0	1	1	6	1	70
	2	1	2	11	17	0	5	1	2	1	0	2	0	2	3	47
	3	3	4	16	27	2	6	0	3	1	0	1	1	2	0	66
	4	5	6	25	37	5	1	0	2	4	0	0	0	2	4	91
	5	9	4	25	25	2	5	6	0	2	4	0	1	3	1	87
	6	5	3	21	27	4	1	3	2	8	2	0	0	2	1	79
	Most 7	8	8	22	19	6	3	2	1	7	0	0	1	2	3	82
	Interested 8	18	18	42	16	3	3	4	1	15	2	2	0	4	2	136
Social Sciences	No Interest 1	3	5	12	7	4	5	1	1	5	0	1	0	8	2	54
	2	1	1	7	3	1	3	1	1	5	1	0	0	2	0	26
	3	5	3	12	10	2	2	1	4	5	0	0	0	1	1	46
	4	9	8	19	15	2	4	4	1	6	3	0	0	2	1	74
	5	10	7	25	32	9	1	5	2	6	0	0	3	2	3	105
	6	8	8	19	37	1	6	1	0	3	1	1	1	2	1	89
	Most 7	4	7	29	37	3	2	0	2	7	0	2	0	2	3	98
	Interested 8	6	2	65	65	2	4	3	0	1	0	1	1	3	3	156
Humanities	No Interest 1	4	2	14	12	2	2	1	0	5	1	0	0	5	1	49
	2	1	1	13	3	1	0	1	1	4	0	0	0	1	1	27
	3	7	5	16	4	1	1	4	3	4	2	0	0	3	4	54
	4	14	9	27	22	1	2	2	3	10	0	1	1	3	0	95
	5	6	6	26	30	4	1	4	2	6	0	1	0	5	0	91
	6	4	7	25	41	2	0	1	1	4	0	1	1	3	3	89
	Most 7	3	8	25	33	6	4	1	1	2	2	2	3	1	2	93
	Interested 8	6	6	21	47	13	20	2	0	2	0	2	2	0	2	123
Mathematics	No Interest 1	4	5	27	49	5	8	0	0	2	1	0	0	7	2	110
	2	0	7	18	24	3	5	0	0	2	0	2	2	1	3	67
	3	5	2	18	26	2	6	0	0	2	0	0	0	0	3	64
	4	4	4	22	15	4	4	1	0	0	0	0	1	2	2	59
	5	6	7	26	17	3	0	0	0	3	0	2	0	3	2	69
	6	10	4	19	22	3	1	3	1	7	0	0	0	2	0	72
	Most 7	11	6	23	14	0	1	5	3	14	3	1	0	1	1	83
	Interested 8	6	11	18	19	4	2	9	8	10	1	0	3	4	1	96
Statistics	No Interest 1	18	15	72	115	14	19	2	1	8	3	1	4	13	8	293
	2	4	4	17	9	3	2	0	0	2	0	2	0	2	0	45
	3	4	3	16	9	4	1	1	0	2	1	0	0	0	0	41
	4	7	5	14	15	1	0	0	1	10	0	0	0	1	0	44
	5	2	6	11	4	0	0	3	3	3	0	1	0	2	0	35
	6	3	5	8	4	1	0	8	1	4	0	0	0	2	0	36
	Most 7	1	0	10	3	0	1	1	1	4	0	0	0	0	0	21
	Interested 8	0	0	0	1	0	0	0	0	0	0	1	0	0	1	3

* The Grand Total = 781 Respondents = 100.00%
Males Total = 400 " = 51.20%
Females Total = 381 " = 48.80%
M refers to Male
F refers to Female

TABLE 4: RATING OF THE QUALITY OF TRAINING IN HIGH SCHOOL AS PERCEIVED BY THE RESPONDENT BY MAJOR FIELD OF STUDY IN COLLEGE

Rating of High School Training in the Fields of Study Stated Below		MAJOR FIELD OF STUDY IN COLLEGE: Natural Sciences M	Natural Sciences F	Social Sciences M	Social Sciences F	Humanities M	Humanities F	Mathematics M	Mathematics F	Physical Sciences M	Physical Sciences F	English M	English F	Other M	Other F	Total
Natural Sciences	Very Poor 1	2	1	18	14	3	1	3	0	0	0	0	0	1	0	39
	2	1	0	11	12	1	6	0	1	0	0	1	0	0	2	35
	3	5	6	22	10	0	1	0	0	5	0	0	0	4	1	54
	4	5	9	21	38	0	6	3	0	2	2	0	0	2	1	89
	5	6	8	32	40	10	4	2	2	9	3	1	2	5	3	127
	6	11	7	34	47	7	5	3	3	9	1	1	0	3	4	135
	7	11	6	33	26	6	2	3	3	10	0	2	3	4	3	112
	Very Good 8	16	17	33	40	3	6	5	3	8	3	1	2	2	4	143
Social Sciences	Very Poor 1	5	3	22	26	2	0	2	0	4	0	0	0	3	2	69
	2	1	1	16	14	0	0	1	0	1	0	1	1	0	0	36
	3	7	5	20	21	2	2	1	1	4	0	0	0	4	0	67
	4	6	8	21	20	6	4	4	1	8	2	0	1	2	2	85
	5	13	9	34	49	7	4	3	1	5	1	2	1	6	3	138
	6	11	4	43	35	5	7	2	3	7	3	0	3	2	3	128
	7	4	3	22	25	5	4	3	1	7	2	3	2	2	2	85
	Very Good 8	6	11	20	22	2	5	1	2	5	0	0	0	1	3	78
Humanities	Very Poor 1	2	5	23	25	2	1	1	0	2	0	0	0	2	0	63
	2	3	1	9	14	0	0	0	0	1	0	0	1	0	1	30
	3	2	2	25	14	4	1	3	1	5	1	0	0	3	0	61
	4	6	10	28	27	0	3	4	0	6	2	0	0	5	2	93
	5	19	8	25	49	6	2	3	4	10	0	2	1	5	3	137
	6	9	5	35	29	6	6	3	4	9	0	2	1	5	3	117
	7	7	9	20	21	6	3	2	1	5	3	3	4	0	2	86
	Very Good 8	5	8	19	31	5	14	1	1	2	0	0	0	1	2	89
Mathematics	Very Poor 1	3	3	15	20	2	3	1	0	1	0	0	0	0	1	49
	2	1	1	5	10	0	2	0	0	1	0	0	0	0	1	21
	3	5	5	13	17	3	4	1	0	3	1	0	0	2	2	56
	4	3	10	18	26	2	4	2	0	1	2	1	0	7	0	76
	5	4	6	36	36	6	3	2	0	7	1	3	1	3	1	109
	6	10	3	28	38	5	4	4	1	12	0	2	1	3	3	114
	7	12	11	31	34	7	2	2	4	7	1	1	2	2	5	121
	Very Good 8	19	14	49	44	6	9	7	7	11	3	0	4	5	4	182
Statistics	Very Poor 1	12	5	55	71	5	7	3	1	8	0	1	1	7	3	179
	2	10	1	16	16	0	2	0	0	3	0	0	1	1	1	51
	3	5	2	14	6	2	1	2	0	3	0	1	0	1	1	38
	4	4	5	12	11	2	2	1	1	6	6	1	0	3	0	54
	5	2	6	11	11	3	0	1	4	5	0	2	0	3	1	49
	6	5	2	10	6	3	1	3	0	1	0	0	0	1	1	33
	7	1	1	6	3	0	0	1	0	4	0	0	1	0	0	17
	Very Good 8	3	2	6	7	1	3	2	0	3	0	0	0	1	1	29

*The Grand Total = 781 Respondents = 100.00%
Males Total = 400 " = 51.20%
Females Total = 381 " = 48.80%
M refers to Male
F refers to Female

TABLE 5: DEGREE OF LIKING OF HIGH SCHOOL TEACHERS OF THE DIFFERENT FIELDS AS PERCEIVED BY THE RESPONDENT BY MAJOR FIELD OF STUDY IN COLLEGE

Degree of Liking of High School Teachers in the Areas of Study Mentioned below			MAJOR FIELD OF STUDY IN COLLEGE														
			Natural Sciences		Social Sciences		Humanities		Mathematics		Physical Sciences		English		Other		Total
			M	F	M	F	M	F	M	F	M	F	M	F	M	F	
Natural Sciences	Disliked	1	1	5	21	15	2	6	2	0	3	0	0	1	4	2	62
		2	1	1	18	15	4	0	1	2	0	0	1	1	2	1	47
		3	2	3	14	15	1	2	0	1	0	0	0	1	1	0	40
		4	5	8	24	40	2	7	4	2	4	1	0	2	3	3	105
		5	10	8	41	38	6	4	5	0	5	1	2	0	9	3	132
		6	11	11	35	39	4	6	2	0	11	2	1	0	1	2	125
	Liked	7	14	6	30	27	9	2	3	2	9	1	0	0	2	5	110
	Very much	8	12	8	22	33	2	5	2	5	10	4	3	2	2	2	112
Social Sciences	Disliked	1	4	1	17	16	1	0	3	0	4	0	0	0	3	1	50
		2	1	3	13	8	1	2	0	1	2	0	1	0	3	0	35
		3	6	3	11	14	2	2	0	2	2	0	0	0	0	0	42
		4	3	8	22	31	4	4	5	1	9	1	1	2	3	2	96
		5	12	7	36	34	10	1	5	2	6	1	3	2	6	2	136
		6	11	5	38	25	3	4	2	1	6	4	0	0	2	2	103
	Liked	7	4	5	25	37	4	3	1	2	7	1	1	3	2	3	48
	Very much	8	9	9	30	33	1	10	1	1	3	1	1	0	1	3	103
Humanities	Disliked	1	4	1	13	14	2	0	2	0	4	1	0	0	1	0	42
		2	3	4	11	3	3	0	0	0	1	0	0	0	1	1	27
		3	4	3	21	16	0	2	2	2	2	1	0	0	2	1	56
		4	4	7	23	28	1	3	3	1	10	2	0	1	4	2	89
		5	11	11	35	35	5	3	3	2	5	0	3	0	5	2	120
		6	8	9	36	37	5	4	3	3	8	0	1	1	3	1	119
	Liked	7	9	7	21	29	6	5	1	0	6	2	2	3	1	6	98
	Very much	8	6	3	23	33	4	13	3	2	3	1	1	2	3	1	99
Mathematics	Disliked	1	4	5	19	26	2	7	0	2	2	2	0	0	3	1	70
		2	1	3	17	13	3	1	0	0	0	0	0	1	2	1	42
		3	4	1	13	20	2	1	2	0	1	0	1	0	1	1	47
		4	6	12	28	29	4	5	2	0	6	0	2	1	4	5	104
		5	8	5	30	37	7	3	0	1	5	1	3	1	5	1	107
		6	12	5	34	29	3	3	3	3	11	0	1	2	1	1	108
	Liked	7	9	7	29	34	6	3	7	3	8	0	0	1	3	4	114
	Very much	8	12	13	25	36	4	9	5	5	10	5	0	2	6	3	135
Statistics	Disliked	1	8	3	28	26	3	4	2	1	3	1	1	0	5	1	85
		2	2	3	15	9	1	1	0	1	2	1	0	0	1	0	36
		3	5	1	14	6	1	1	0	0	3	1	0	0	1	1	34
		4	1	4	11	11	2	1	1	0	5	1	1	0	4	1	43
		5	9	3	16	4	1	1	3	1	4	0	0	0	3	0	46
		6	1	2	9	8	1	0	4	1	3	0	1	0	1	0	31
	Liked	7	3	1	7	2	0	0	0	0	4	0	1	0	0	1	19
	Very much	8	2	2	5	5	0	1	0	0	2	0	0	0	0	1	18

* The Grand Total = 781 Respondents = 100.00%
Males Total = 400 " = 51.20%
Females Total = 381 " = 48.80%
M refers to Male
F refers to Female

TABLE 6: RATING OF THE SOCIAL SCIENCES AND MATHEMATICS CONTRIBUTIONS TO THE WELL BEING OF THE INDIVIDUAL AS PERCEIVED BY THE RESPONDENTS WHO ARE MAJORING IN SOCIAL SCIENCE IN COLLEGE VS. THOSE MAJORING IN MATHEMATICS

Field of Study	Areas It Contributes To	Ratings of Social Science Students												Ratings of Mathematics Students												Mean	Standard Deviation
		MALE %						FEMALE %						MALE %						FEMALE %							
		Very low					Very high	Very low					Very high	Very low					Very high	Very low					Very high		
		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	N	SD
Social Science	Social Good Lead Better Life	3.5	1.0	7.6	22.2	26.3	39.4	2.8	2.3	7.9	21.4	23.3	42.3	5.6	5.6	11.1	27.8	22.2	27.8	0.0	0.0	7.7	15.4	53.8	23.1	4.7	1.3
	Useful in job	5.1	3.6	10.2	26.5	26.5	28.1	5.1	6.0	11.2	22.8	25.1	29.8	5.6	16.7	22.2	16.7	22.2	16.7	0.0	0.0	7.7	46.2	38.5	7.7	4.3	1.4
	Understanding public issues	9.7	7.1	8.7	23.5	27.6	23.5	5.6	3.7	10.7	24.7	27.0	28.4	5.6	16.7	16.7	33.3	11.1	16.7	0.0	0.0	25.0	41.7	25.0	8.3	4.2	1.4
	Make more money	2.6	5.1	6.2	14.4	34.4	37.4	2.3	1.9	6.5	19.6	30.8	38.8	0.0	5.6	16.7	38.9	33.3	5.6	0.0	0.0	16.7	16.7	50.0	16.7	4.8	1.2
	Developing Personality	14.0	14.5	25.9	24.9	14.0	6.7	12.7	8.5	23.9	28.6	13.1	13.1	22.2	16.7	22.2	27.8	5.6	5.6	0.0	15.4	15.4	30.8	23.1	15.4	3.4	1.5
	Developing Mental Ability	6.2	6.7	12.9	19.1	33.5	21.6	6.4	8.7	12.8	18.3	25.7	27.5	5.9	29.4	11.8	29.4	23.5	0.0	0.0	8.3	25.0	25.0	33.3	8.3	4.2	1.5
	Developing Personal rationality and logic	4.1	9.3	14.5	23.8	24.9	23.2	4.6	6.5	15.3	21.8	22.7	27.3	5.9	11.8	11.8	47.1	17.6	5.9	0.0	0.0	30.8	23.1	38.5	7.7	4.2	1.5
		3.6	4.1	10.3	24.2	29.9	27.8	6.0	5.1	12.1	23.7	26.5	26.5	17.6	5.9	11.8	41.2	17.6	5.9	0.0	7.7	23.1	30.8	30.8	7.7	4.4	1.4
Mathematics	Social Good Lead Better Life	30.0	18.9	16.8	13.7	11.1	9.5	29.2	22.6	18.9	15.6	9.4	4.2	16.7	38.9	11.1	22.2	0.0	11.1	0.0	15.4	30.8	30.8	7.7	15.4	2.9	1.6
	Useful in Job	30.4	18.8	18.8	14.1	12.6	5.2	30.3	23.2	20.4	12.8	9.5	3.8	5.6	33.3	22.2	22.2	5.6	11.1	7.7	15.4	23.1	38.5	7.7	7.7	2.8	1.5
	Understanding public issues	9.0	6.9	10.1	20.2	30.9	22.9	11.5	10.5	13.9	18.7	21.1	24.4	0.0	5.6	22.2	5.6	27.8	38.9	0.0	0.0	7.7	0.0	23.1	69.2	4.3	1.6
	Make more money	34.9	19.8	20.8	14.6	6.3	3.6	29.5	21.0	26.7	11.9	7.6	3.3	11.1	44.4	27.8	16.7	0.0	0.0	7.7	23.1	61.5	7.7	0.0	0.0	2.6	1.4
	Developing Personality	14.8	5.8	13.8	21.7	27.0	16.9	16.2	9.0	19.0	11.9	22.9	21.0	0.0	11.1	0.0	38.9	27.8	22.2	0.0	7.7	7.7	0.0	30.8	53.8	3.9	1.7
	Developing Mental Ability	33.3	24.9	20.1	13.2	5.3	3.2	43.8	22.1	19.2	8.2	3.4	3.4	27.8	27.8	33.3	11.1	0.0	0.0	15.4	15.4	38.5	30.8	0.0	0.0	2.4	1.4
	Developing Personal rationality and logic	10.5	6.8	8.4	16.2	25.7	32.5	15.5	9.9	10.8	14.6	18.8	30.5	5.6	11.1	16.7	0.0	38.9	27.8	7.7	0.0	7.7	7.7	30.8	46.2	4.3	1.7
		9.9	5.8	12.6	16.2	19.9	35.6	16.7	11.0	10.5	14.8	19.0	28.1	0.0	11.1	22.2	5.6	22.2	38.9	7.7	0.0	7.7	7.7	30.8	46.2	4.2	1.7

Grand Total = 781 Respondents = 100.00%
Males Total = 400 " = 51.20%
Females Total = 381 " = 48.80%

DISTRIBUTED LAG MODELS AND THE EFFECTS OF SOCIOECONOMIC CHANGE ON FERTILITY

William Ray Arney
University of Colorado

Interest in the relationship between socioeconomic and demographic change has developed because of the widespread attention given the theory of demographic transition (Thompson, 1929, 1944; Notestein, 1945). Briefly, the theory states that there are three stages in the history of a population: (1) the population is stable and a regime of high birth and death rates prevails, (2) the population undergoes a transition in which death rates decline followed by a decline in fertility, and (3) the population reaches a fairly stable state which is characterized by low birth and death rates. The relationship between socioeconomic and demographic change becomes most important at the second or transitional stage of development. It has been suggested that technology to control mortality is applied more ubiquitously and rapidly under conditions of high urbanization and industrialization. This, in connection with the high fertility rates, causes rapid population growth. Fertility responds with a downward trend, according to speculation, because high fertility handicaps the population "in their effort to take advantage of the opportunities being provided by the emerging economy" (Davis, 1963: 352).

Students of population dynamics are quick to point out that the theory of demographic transition initially was simply an empirical generalization with no real rationale (Gutman, 1960). However, in more recent work we find several schemes which attempt to provide rationales for the observations. For example, Cowgill begins with the Malthusian supposition that a population will grow at a geometric rate until it approaches a size which begins to affect resource and space availability. He believes that most populations "at any given time" have reached this point and have achieved "a condition of equilibrium characterized by a relatively stationary population" (1963: 271). His refinement of transition theory is contained in the assertion that technological advances increase the "carrying capacity of the environment" (presumably a function of resource and space availability) which leads to a period of population increase. He explicitly states:

> "Under conditions of industrialization, given the technology of birth and death control, there is a marked tendency for the technology of death control to be applied earlier and more extensively resulting in rapid population growth ..." (p. 272).

Generally congruent with Cowgill's propositions is the theory of economic and demographic interdependence developed by Frederiksen (1969). His model posits complex interaction among variables such as technological and socioeconomic development, production, "levels of nutrition, sanitation, health services, etc.," mortality, and fertility. With considerable simplification his theory can be stated as follows. With increasing technological development comes increasing health care distribution and implementation which causes a decrease in mortality. Decreased mortality helps decrease fertility by increasing survival probabilities of offspring and thereby decreasing the need for large families. This model views a decline in mortality, therefore, not as a cause of the population problem but as a necessary factor for the solution of it.

On the basis of these two important theoretical contributions we can identify one major factor which should be included in an explanation of changes in fertility. Urbanization brings about changes which are more conducive to mortality decline which can lead to a reduction of fertility. Therefore, at the most general level, one might expect urbanization and fertility to be inversely related.

A second set of factors which should theoretically influence fertility is derived from economic theories of fertility behavior. Generally, these variables influence fertility through mechanisms regulated by opportunity costs (Gronau, 1973; Mincer, 1963). It is thought that as the social status of women increases (in terms of education, economic earning power, and so forth) the costs of having children increases. That is, having children removes the opportunity to earn more money, or in some way other than raising children, to make use of the woman's time. Thus, variables related to the social and economic position of women in a society become an important factor in the analysis of fertility.

Unfortunately, data related to the employment of women of the form required for this study (long time series data) were unavailable. Therefore, it was decided to use a proxy which was though to be indicative of overall opportunity costs -- unemployment. It was hypothesized that as unemployment increased, opportunity costs associated with having children would decrease. As opportunity costs decrease, fertility should increase; so a direct relationship between unemployment and fertility was expected.

A third factor assumed to be affecting fertility was religious composition of a population. Heer and Boynton (1970) found that the second highest correlation in their study of data for counties in the United States was between fertility and the proportion of Roman Catholics in the diocese or archdiocese in which the county was located. Although the reproductive ideals and behavior of Roman Catholics seems to be changing it was thought that, particularly for the time period which the data for this study covers -- 1919 to 1967 -- an important variable to include in the analysis was the percentage of the population who were Roman Catholic. A direct relationship between the percentage of Roman Catholics in the population and fertility was expected.

PROCEDURE

Considerable effort has been devoted to the study of the effects of socioeconomic change on fertility. Previously, most studies have used cross-sectional data and various forms of regression analysis. Although this has the merits of data availability (at least more so than the time series data that will be used here) and extensive technical development, the cross-sectional approach

is inadequate for two reasons. First, in order to make inferences concerning developmental processes it is necessary to infer a developmental dimension into the cross-sectional data or as Goldscheider (1971: 85) puts it, one must exercise an "evolutionary bias". He notes,

> "The foundation of prediction and control ... assumes that currently nonmodern nations will follow similar patterns of development experienced by currently modern nations, and that specific relationships between social, economic, political, and cultural variables on one hand and population variables on the other will be the same in the modernization process of developing societies as they were in the historical experience of developed societies."

The present study avoids this problem by examining the development of a single country as that development is reflected in historical time series. That is, the time dimension is considered explicitly and no time dependent variable need be inferred into the data or the results of the analysis.

Second, regression analysis requires a rather strict parametric structure. Heise (1969) calls this the problem of specification. The procedures used here to estimate models of fertility circumvents the specification problem at least to a degree. Also, depending on data collection procedures, the lag structure of a regression model can be restricted. For example, if data on divorces and marriages were collected from a number of "most recent censuses" the regression model which would result would relate the variables at only one point in time. This is clearly unrealistic since one would logically expect divorce rates to rise a certain number of years after an increase in marriages. In other words, divorces should lag marriages and a model of the relationship between these variables should reflect this. Even more realistically, since "length of marriage to divorce" is distributed over a number of years one should expect an increase in marriages in year x to cause an increase in divorces in years x, x+1, x+2, ... A model to describe this relationship is called a distributed lag model. In fact, Carlsson (1970) has demonstrated the utility of the distributed lag concept with his model of the relationship between fertility and marriages in nineteenth century Sweden.

In this study, time series data for the United States is used and spectral analysis is employed to estimate a distributed lag relationship between fertility on the one hand, and urbanization, unemployment, and Roman Catholic population on the other. Yearly data on each of the variables was collected from published sources. Urbanization was measured by the complement of the farm population which has been estimated annually by the U.S. Department of Agriculture. Estimates of the proportion of the population belonging to the Roman Catholic church were obtained from series H-538 of the Historical Statistics of the United States (Washington, D.C.: Government Printing Office, 1960; continuation, 1965) and various issues of the Statistical Abstract of the United States. Unemployment was taken from series D-47 of the Historical Statistics and issues of the Statistical Abstracts.

Births per 1,000 women aged 15-44 years (computed by summing birth rates by age of mother in five year age groups multiplied by five) was chosen as the dependent variable. Estimates of this variable are available on a yearly basis since 1909 from the Bureau of Vital Statistics. This indicator of fertility has the advantage over other measures (e.g., the crude birth rate) that an adjustment has been made for age composition. This is beneficial since there is an interaction between fertility and a population's age distribution.

The use of spectral analysis to estimate distributed lag models is reviewed in detail in other sources (Jenkins and Watts, 1968; Fishman, 1969; Hannan, 1963, 1965, 1967; Mayer and Arney, 1974). Due to space limitations it will be reviewed here with considerable brevity.

If we assume that a discrete time series $\{x_t\}$ is related to the series $\{y_t\}$ then a distributed lag relating the two has the form

$$(1) \qquad y_t = \sum_{i=0}^{\infty} h_i \, x_{t-i} \, .$$

In order to estimate the collection of constants $\{h_i\}$ we must first make an assumption about and place certain restrictions on the series $\{x_t\}$ and $\{y_t\}$. First, it must be assumed that the series are realizations of discrete stochastic processes $\{X_t\}$ and $\{Y_t\}$. Second, in order to use spectral analysis the underlying stochastic processes must be covariance stationary. That is, the covariance structure of the processes must not be dependent on the ordering variable, t, which is usually taken to be time.

Under the above restrictions equation 1 can be considered to be a representation of the system which linearly relates the discrete covariance stationary processes $\{X_t\}$ and $\{Y_t\}$ as

$$(2) \qquad Y_t = \sum_{i=0} h_i \, X_{t-i} + Z_t \, .$$

The Z_t in equation 2 are terms of a discrete white noise process. That is, $\{Z_t\}$ is a discrete stochastic process in which all terms are independent of one another and each of the terms has the same distribution. The set of constants $\{h_i\}$ is called the impulse response function of the system. The problem becomes one of solving equation 2 for the impulse response function $\{h_i\}$.

A theorem due to the mathematician Norbert Weiner provides an initial step toward the solution of this problem. The Weiner-Hopf theorem states that the impulse response function in equation 2 which minimizes the mean square error of the linear prediction must also satisfy the relation

$$(3) \qquad \gamma_{xy}(u) = \sum_{i=0} h_i \, \gamma_{xx}(u-i)$$

where the γ_{xx} and γ_{xy} are the autocovariance function of $\{X_t\}$ and the cross-covariance function of $\{X_t\}$ and $\{Y_t\}$, respectively. Using equation 3 and spectral functions it will be possible to solve for $\{h_i\}$.

The spectrum of the process $\{X_t\}$, denoted $\Gamma_{xx}(f)$, is the Fourier transform of the autocovariance function of $\{X_t\}$. Similarly, the cross-spectrum of the processes $\{X_t\}$ and $\{Y_t\}$, denoted $\Gamma_{xy}(f)$, is the Fourier transform of the cross-covariance function of $\{X_t\}$ and $\{Y_t\}$. If we let H(f) be the Fourier

transform of the impulse response function then, due to the mathematical nature of Fourier transforms the frequency domain representation of equation 3 is

$$\Gamma_{xy}(f) = H(f)\,\Gamma_{xx}(f). \tag{4}$$

H(f) is called the frequency response function of the system. Equation 4 can now be solved for H(f) by

$$H(f) = \frac{\Gamma_{xy}(f)}{\Gamma_{xx}(f)}. \tag{5}$$

Since H(f) is the Fourier transform of the impulse response function, $\{h_i\}$, the impulse response function can be found by taking the inverse transform of H(f).

A system with multiple inputs $\{X_{k,t}, k = 1, 2, \ldots n\}$ and a single output $\{X_{n+1,t}\}$ can be represented by

$$X_{n+1,t} = \sum_{k=1}^{n} \sum_{i=0}^{\infty} h_{k,i}\, X_{k,t-i} + Z_t \tag{6}$$

Following reasoning similar to that for the bivariate can it can be shown that if $G_{n+1}(f)$ is a vector of cross-spectra between each input and the output series, $G_{nn}(f)$ is a square matrix of spectra and cross-spectra among all the inputs, and $H_{n+1}(f)$ is a vector of partial frequency response functions, then

$$H_{n+1}(f) = G_{n+1}(f)\, G_{nn}^{-1}(f) \tag{7}$$

provided the inverse of $G_{nn}(f)$ exists. The partial impulse response functions $\{h_{k,i}, k = 1, 2, \ldots, n\}$ can be found by taking the inverse Fourier transforms of the series of vectors $H_{n+1}(f)$.

There are two major adjustments of the data which must be made before spectral analysis can be used to estimate the coefficients in a model like equation 6. The data must be filtered to approximate covariance stationarity and then the independent series must be aligned with the dependent series to avoid biasing the spectral functions.

Generally, a first or second difference filter is sufficient to remove severe non-stationary components of a time series. A first order difference filter has the form

$$x_t^* = x_t - x_{t-1} \tag{8}$$

where $\{x_t^*\}$ is the filtered series, and a second difference filter is merely the result of applying a first difference filter twice

$$x_t^* = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) \tag{9}$$

$$= x_t - 2x_{t-1} + x_{t-2}.$$

These two filters can be thought of as the discrete analogs of derivatives. A second difference filter was necessary to achieve approximate covariance stationarity in the fertility and Roman Catholic population series. The non-stationarity evident in the urbanization and unemployment series was reduced by a first difference filter. The criterion used to judge approximate stationarity was a noticable reduction in the low frequency variation of the series as shown by the spectrum of the filtered series.

After filtering the data one must align the filtered independent series with the filtered dependent series. This is a requirement of cross-spectral analysis, but it also allows one to at least make an "educated guess" about the structure of the system with which one is working. That is, alignment requires the specification of a lead or lag relationship between two series.

It will be recalled that in order to estimate the cross-spectrum of two processes the cross-covariance function must be estimated first. The estimate of the cross-covariance function used in this paper is

$$\gamma_{xy}(u) = \frac{1}{n-u} \sum_{i=1}^{n-u} (x_i - \bar{x})\,(y_{i+u} - \bar{y}) \tag{10}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the series $\{x_t\}$ and $\{y_t\}$ respectively. The cross-covariance function will peak at a value of u, u', where the covariance between one series and the other series shifted u' units is greatest. If the series are not aligned, i.e., adjusted so that the greatest covariance occurs at zero lags, the estimated spectral functions suffer considerable bias. For example, inspecting the cross-covariance function between the filtered urbanization and fertility series it was found that urbanization lagged fertility by four years. Accordingly the filtered urbanization series was transformed by

$$u_t^{**} = u_{t-4}^{*} \tag{11}$$

and all spectral and cross-spectral functions were estimated for the series $\{u_t^{**}\}$ and $\{f_t^{*}\}$. Similarly the following adjustments of the other independent series were required

$$\text{for unemployment: } v_t^{**} = v_{t-1}^{*} \tag{12}$$

$$\text{for Roman Catholic pop.: } r_t^{**} = r_{t-3}^{*} \tag{13}$$

where a single asterik denotes the filtered series and a double asterik denotes the filtered, aligned series.

RESULTS

Using the procedure outlined above a distributed lag model between the filtered and aligned series was estimated. The partial impulse response functions were obtained by applying the inverse Fourier transformation to the partial frequency response functions. The partial impulse response functions appear in Table I.

Model I

One problem in constructing a multivariate distributed lag model is how to choose the number of terms from each partial impulse response function to include in the model. One solution to this problem involves the inspection of the cross-covariance function between two series. If the cross-covariance function drops off rapidly after p lags, it is probable that p terms of that impulse response function should be included in the model.

Using this technique it was decided that one term of each of the first two partial impulse response functions (urbanization and unemployment) and five terms of the third partial impulse response function should be included in the model.

This resulted in the equation

$$(14)\quad f_t^* = -4.75\, u_t^{**} + .78\, v_t^{**} - .039\, r_t^{**} - 3.92\, r_{t-1}^{**} + 3.95\, r_{t-2}^{**} - .17\, r_{t-3}^{**} - 1.81\, r_{t-4}^{**} \quad .$$

Equation 14 will be called Model I. It explains only 11% of the filtered fertility series suggesting that this model is rather inadequate.

Model II

An alternative method for selecting terms of the partial impulse response functions to enter the model is a computer search to meet some criterion. In this case, variance explained by a linear projection was maximized. The model constructed in this way is

$$(15)\quad f_t^* = -4.75\, u_t^{**} + .78\, v_t^{**} - 1.23\, v_{t-4}^{**} + .13\, v_{t-5}^{**} + .45\, v_{t-6}^{**} - .34\, v_{t-7}^{**} - .039\, r_t^{**} - 3.92\, r_{t-1}^{**} + 3.95\, r_{t-2}^{**} - .17\, r_{t-3}^{**} - 1.81\, r_{t-4}^{**} \cdot$$

This model is certainly a better predictor of fertility since it explains approximately 39% of the variance in the filtered fertility series. One method of evaluating the adequacy of this model is to compare the original spectrum of the filtered fertility series to the spectrum of the residual series. If the residual series were purely random the residual spectrum would be essentially flat (Jenkins and Watts, 1968: 224-225). Figure I shows that considerable flattening has occurred even though the residual spectrum is not completely flat.

Notice that equation 15, Model II, is similar to Model I in that it has one term of the partial impulse response function associated with the urbanization series and five term of the "Roman Catholic population" partial impulse response function. However, Model II differs significantly from Model I with respect to the unemployment coefficients. Terms corresponding to lags of zero, four, five, six, and seven years enter the model. This is especially curious since the terms of the partial impulse response function corresponding to lags of one, two, and three years are certainly not zero.

Is it possible to reconcile this finding? It is possible that the secondary delay is due to some form of recursion, i.e., unemployment exerting an influence on fertility indirectly through another variable. To check this speculation the cross-covariance functions between the unaligned unemployment series and the other two independent series were examined. It was found that a very strong relationship existed between urbanization and unemployment lagged one year. Recall that urbanization required an adjustment of four years to achieve proper alignment with fertility. These four years together with the lag of one year of unemployment behind urbanization suggests an indirect effect of unemployment on fertility occurring with a lag of five years through the intervening variable urbanization. This indirect effect occurs at approximately the same lag as suggested by the structure of equation 15. In other words, a system of the form in Figure II seems to be operative here.

The total effect of unemployment on fertility can be found by adding the convolution of the distributed lag relating unemployment to urbanization and the distributed lag relating urbanization to fertility with the direct effect of unemployment on fertility. By methods described above it was found that the best linear predictor of urbanization based on unemployment was

$$(16)\quad u_t^* = -.1615\, v_{t-1}^*$$

This, when convoluted with the term of the partial impulse response function associated with the urbanization series which was included in Model II yields an equation which is quite similar to Model II.

$$(17)\quad f_t^* = -4.75\, u_t^{**} + .78\, v_t^{**} - .768\, v_{t-4}^{**} - .039\, r_t^{**} - 3.92\, r_{t-1}^{**} + 3.95\, r_{t-2}^{**} - .17\, r_{t-3}^{**} - 1.81\, r_{t-4}^{**} \cdot$$

As can also be seen the magnitude of the coefficient for the v_{t-4}^{**} term is similar to the same coefficient in Model II. This secondary analysis lends support to the above speculation concerning the nature of the operative system.

DISCUSSION

If we remove the alignment and formulate equation 15 in terms of differencing operators Δ we obtain

$$(18)\quad \Delta f_t = \Delta f_{t-1} - 4.75\, \Delta u_{t-4} + .78\, \Delta v_{t-1} - 1.23\, \Delta v_{t-5} + .131\, \Delta v_{t-6} + .45\, \Delta v_{t-7} - .039\, \Delta r_{t-3} - 3.89\, \Delta r_{t-4} + 7.88\, \Delta r_{t-5} - 4.12\, \Delta r_{t-6} - 1.64\, \Delta r_{t-7} + 1.81\, \Delta r_{t-8}$$

As can be seen the most immediate effects of unemployment and urbanization are in the expected directions. Urbanization has a strong negative effect on fertility change after a lag of four years. The initial effect of unemployment on changes in fertility is positive and occurs at a lag of one year lending support to the opportunity cost argument concerning the influence of economic change on fertility. The longer term effects of unemployment on fertility are decidedly negative which suggests that in the long run income effects come into play. Perhaps the most confusing distributed lag relationship is that between Roman Catholic population and fertility. The strongest influence is positive and occurs at a lag of five years. However, the positive effect is offset somewhat by the negative effects at lags of three, four, six, and seven years.

In future papers the analysis performed here will be extended in several ways. First, other than substantive uses there is a question concerning possible uses of the types of models developed here. It is possible that such models could be used for purposes of projection and prediction of future trends in fertility on the basis of socioeconomic change. This will be explored. Second, the present model is being extended to include constructed recursive effects. In other words, instead of detecting recursion as was done in the present work recursive effects of variables will be "built in" by the investigator in theoretically meaningful ways. For example, it is thought that urbanization has an indirect effect on fertility through its influence on health care distribution and that variable's subsequent effect on infant mortality. This, then, leads to a reduction in fertility.

Much work remains to be done in this area. The present study is preliminary at best. We proceeded by effectively ignoring problems of data consistency over time and other methodological difficulties. Essentially, we were trying to answer the question of whether this approach can be useful at all in sociology. Inasmuch as it provides a new framework (with its own theoretical implications) for the analysis of data of a form not typically used by sociologists, the answer to the question is that the method appears useful, but further research may serve to qualify this initial response.

REFERENCES

Carlsson, Gosta.
1970 "Nineteenth-century fertility oscillations." Population Studies 24: 413-422.

Cowgill, Donald O.
1963 "Transition theory as general population theory." Social Forces 41: 270-274.

Davis, Kingsley.
1963 "The theory of change and response in modern demographic history." Population Index 29: 345-365.

Fishman, George S.
1969 Spectral Methods in Econometrics. Cambridge, Mass.: Harvard University Press.

Frederiksen, Harald.
1969 "Feedbacks in economic and demographic transition." Science 166: 837-847.

Goldscheider, Calvin.
1971 Population, Moderization, and Social Structure. Boston: Little, Brown, and Company.

Gronau, Reuben.
1973 "The effect of children on the housewife's value of time." Journal of Political Economy 81: S168-S199.

Gutman, Robert.
1960 "In defense of population theory." American Sociological Review 25: 325-333.

Hannan, E. J.
1963 "Regression for time series." Pp. 17-37 in M. Rosenblatt (ed.), Proceedings of the Syposium on Time Series Analysis. New York: John Wiley.
1965 "The estimation of relationships involving distributed lags." Econometrica 33: 206-224.
1967 "The estimation of a lagged regression relation." Biometrika 54: 409-418.

Heer, David M., and John W. Boynton.
1970 "A multivariate regression analysis of differences in fertility of United States counties." Social Biology 17: 180-194.

Heise, David R.
1969 "Problems in path analysis and causal inference." Pp. 38-73 in E. F. Borgatta (ed.), Sociological Methodology 1969. San Francisco: Jossey-Bass.

Jenkins, Gwilym M., and Donald G. Watts.
1968 Spectral Analysis and Its Applications. San Francisco: Holden-Day.

Mayer, Thomas F., and William Ray Arney.
1974 "Spectral analysis and the study of social change." Pp. 309-355 in H. L. Costner (ed.), Sociological Methodology 1973. San Francisco: Jossey-Bass.

Mincer, Jacob.
1963 "Market prices, opportunity costs, and income effects." Pp. 67-82 in C. Christ, et. al (eds.), Measurement in Economics. Stanford: Stanford University Press.

Notestein, Frank W.
1945 "Population - the long view." Pp. 36-57 in T. W. Schultz (ed.), Food for the World. Chicago: University of Chicago Press.

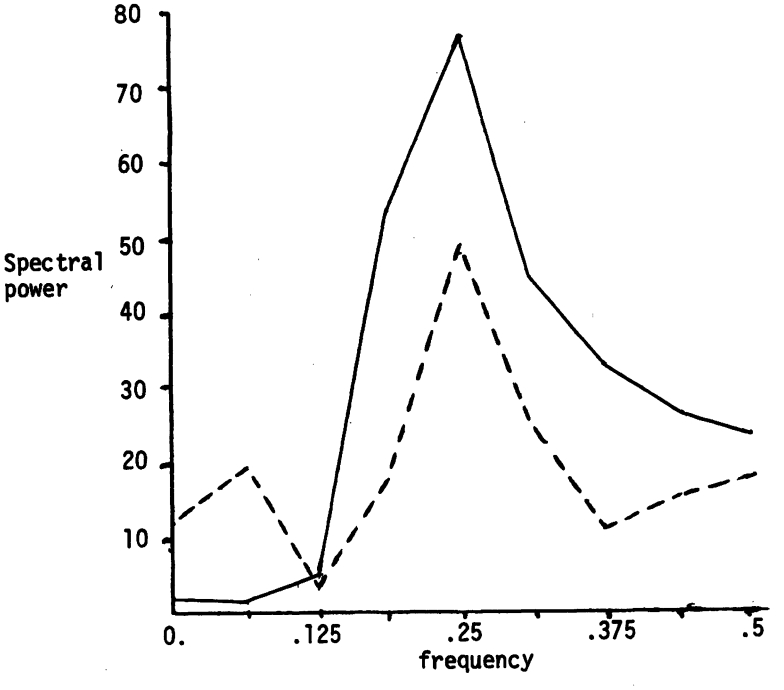
Thompson, Warren S.
1929 "Population." American Journal of Sociology 34: 959-975.
1944 Plenty of People. Lancaster, Pa.: Jacques Cattell Press.

TABLE I

Partial impulse response functions for a model of fertility based on the filtered, aligned urbanization, unemployment, and Roman Catholic population series.

lags	partial impulse response functions		
	urbanization	unemployment	Roman Catholic population
0	-4.754	.782	-.039
1	-3.845	-1.630	-3.929
2	2.226	.248	3.956
3	2.620	1.159	-.172
4	-1.154	-1.236	-1.815
5	-5.039	.131	1.954
6	3.321	.453	-1.183
7	.858	-.340	-1.370

Figure I: Original spectrum of the filtered fertility series and the residual spectrum derived from Model II.



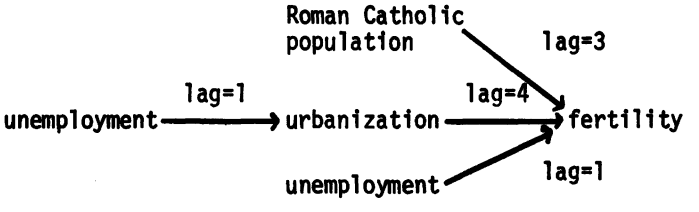


Figure II: Relationships among filtered urbanization, unemployment, Roman Catholic population, and fertility series.

AN UNBIASED RESPONSE MODEL FOR ANALYSIS OF CATEGORICAL DATA

George E. Battese and Wayne A. Fuller, Iowa State University

1. INTRODUCTION

In this paper we consider the estimation of multinomial proportions when sample determinations contain classification or response errors. It is well known that if each individual in a simple random sample from a population is classified "correctly" then the sample proportions are unbiased estimators for the corresponding population proportions. However, in the presence of response errors the sample proportions are not necessarily unbiased for the population proportions. Given the response probabilities we present expressions for the expectations, variances and covariances for the sample proportions. In particular, we consider a response model that has the property that the sample proportions are unbiased estimators for the population proportions.

The effects of response errors on methods of analysis of categorical data have been considered by Bross (1954), Mote and Anderson (1965), Assakul and Proctor (1967), Koch (1969), McCarthy (1972), and others. Different response models have been discussed by Giesbrecht (1967), Bershad (1967), Koch (1968), Huang (1972), and Battese (1973).

In our discussion we assume that a simple random sample of size n is selected without replacement from a finite population of N individuals. Each individual is classified into one of r mutually exclusive and exhaustive classes. The sample classification is assumed to depend on the true class through parameters β_{ij}, i, j = 1, 2, ..., r, where $\sum_{j=1}^{r} \beta_{ij} = 1$, for all i = 1, 2, ..., r. The parameter, β_{ij}, is the probability that a randomly selected individual belonging to the i-th class is classified into the j-th class.

The proportion of the sample that is classified in class i is denoted by

$$p_i = \frac{1}{n} \sum_{k=1}^{N} \delta_k \gamma_k^i \qquad (1.1)$$

where $\delta_k = 1$ if the k-th population individual is in the sample;

$= 0$ otherwise; and

$\gamma_k^i = 1$ if the k-th population individual is classified in class i;

$= 0$ otherwise.

It is readily verified that under these response hypotheses

$$E(p_i) = \sum_{m=1}^{r} P_m \beta_{mi} = \bar{P}_i \qquad (1.2)$$

$$\mathrm{Var}(p_i) = \frac{(N-n)}{n(N-1)} \bar{P}_i(1-\bar{P}_i) + \frac{(n-1)}{n(N-1)} \sum_{m=1}^{r} P_m \beta_{mi}(1-\beta_{mi}) \qquad (1.3)$$

$$\mathrm{Cov}(p_i, p_j) = -\frac{(N-n)}{n(N-1)} \bar{P}_i \bar{P}_j - \frac{(n-1)}{n(N-1)} \sum_{m=1}^{r} P_m \beta_{mi} \beta_{mj} \qquad (1.4)$$

where P_i, i = 1, 2, ..., r, denote the proportions of the population in the different classes.

The expression of (1.2) shows that the presence of response errors can result in the sample proportions being biased estimators of the true proportions.

2. AN UNBIASED RESPONSE MODEL

If the sample responses are such that the true response is reported a fraction, α, of the time and for the remaining fraction, $(1-\alpha)$, of the time the response is given with probabilities proportional to the population parameters P_i, i = 1, 2, ..., r, then the response probabilities are

$$\beta_{ii} = \alpha + (1-\alpha)P_i, \quad i = 1, 2, \ldots, r \qquad (2.1)$$

$$\beta_{ij} = (1-\alpha)P_j, \quad i \neq j;\ i, j = 1, 2, \ldots, r \qquad (2.2)$$

where α is a constant in the interval [0, 1].

For this response model the sample proportion for any given class unbiasedly estimates the true proportion belonging to that class. That is, $\sum_{m=1}^{r} P_m \beta_{mi} = P_i$. The variances and covariances of the sample proportions for this response model are

$$\mathrm{Var}(p_i) = \frac{(N-n)}{n(N-1)} P_i(1-P_i) + \frac{(n-1)}{n(N-1)} (1-\alpha^2)P_i(1-P_i) \qquad (2.3)$$

$$\mathrm{Cov}(p_i, p_j) = -\frac{(N-n)}{n(N-1)} P_i P_j - \frac{(n-1)}{n(N-1)} (1-\alpha^2)P_i P_j . \qquad (2.4)$$

It is obvious that the response parameter, α, in the unbiased response model (2.1, 2.2) cannot be estimated without repeated responses from sample individuals. We consider the

estimation of α and the (r-1) independent proportions, P_i, i = 1, 2, ..., r-1, from two independent responses on each sample individual. These two responses are assumed to be those obtained in "Trial 1" (an original interview) and "Trial 2" (a reinterview) of a survey. The proportion of the sample which responds in class i at Trial 1 and class j at Trial 2 is denoted by

$$p_{ij} = \frac{1}{n} \sum_{k=1}^{N} \delta_k \gamma_k^{ij} \qquad (2.5)$$

where γ_k^{ij} = 1 if the k-th individual is classified in class i at Trial 1 and class j at Trial 2;

= 0 otherwise.

Given that the Trial-1 and Trial-2 responses are independent, it can be verified that, for a general response model, the expectations, variances and covariances of the two-trial proportions are

$$E(p_{ij}) = \sum_{m=1}^{r} P_m \beta_{mi} \beta_{mj} \equiv P_{ij} \qquad (2.6)$$

$$\mathrm{Var}(p_{ij}) = \frac{(N-n)}{n(N-1)} P_{ij}(1-P_{ij}) + \frac{(n-1)}{n(N-1)} \sum_{m=1}^{r} P_m \beta_{mi} \beta_{mj} (1-\beta_{mi}\beta_{mj}) \qquad (2.7)$$

$$\mathrm{Cov}(p_{ij}, p_{i'j'}) = -\frac{(N-n)}{n(N-1)} P_{ij} P_{i'j'} - \frac{(n-1)}{n(N-1)} \sum_{m=1}^{r} P_m \beta_{mi} \beta_{mj} \beta_{mi'} \beta_{mj'}, \quad i \neq i' \text{ or } j \neq j' . \qquad (2.8)$$

Under the assumption of independence of the Trial-1 and Trial-2 responses, it is evident from (2.6) that the expected proportions P_{ij} and P_{ji} are equal. This implies that, at most, there exists $\frac{1}{2}(r+2)(r-1)$ independent expected proportions, P_{ij}. Further, the result of (2.6) suggests that large differences in the observed proportions p_{ij} and p_{ji}, $i \neq j$, may indicate lack of independence of the Trial-1 and Trial-2 survey responses. An approximate test for "lack of symmetry of the expected two-trial proportions," or equivalently, "lack of independent classifications in two trials" is obtained with the statistic

$$X_S^2 = \sum_{i<j}^{r} \sum_{j}^{r} (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji}) \qquad (2.9)$$

where n_{ij} denotes the number of the n sample individuals that are classified in the i-th class on Trial 1 and the j-th class on Trial 2 ($n_{ij} = np_{ij}$). This statistic converges to a central chi-square random variable with $\frac{1}{2}(r^2 - r)$ degrees of freedom under the hypothesis that $P_{ij} = P_{ji}$ for all i and j.

The expectations of the two-trial proportions for the unbiased response model (2.1, 2.2) are

$$P_{ii} = \alpha^2 P_i + (1-\alpha^2) P_i^2 \qquad (2.10)$$

and

$$P_{ij} = (1-\alpha^2) P_i P_j , \quad i \neq j . \qquad (2.11)$$

The likelihood function for the Trial-1 and Trial-2 responses is that of the multinomial distribution with r^2 classes having probabilities, P_{ij}, i, j = 1, 2, ..., r, defined by (2.10) and (2.11). The maximum likelihood estimators for the independent parameters, P_i, i = 1, 2, ..., r-1, and α, are not readily obtainable from the likelihood equations [see Battese (1973)]. The Gauss-Newton estimators are, however, more easily obtained.

Given that the vector of the r^2-1 independent two-trial proportions is expressed by

$$Y = (p_{11}, p_{12}, \cdots, p_{1r}, \cdots, p_{r1}, p_{r2}, \cdots, p_{r,r-1})' \qquad (2.12)$$

we write the model

$$Y = P(\theta) + e \qquad (2.13)$$

where $P(\theta)$ denotes the vector of expected values of the sample proportions expressed as functions of the vector of independent parameters, θ; and e denotes the vector of the deviations of the observed proportions from the expected proportions. By expressing $P(\theta)$ as a Taylor expansion about an initial estimate for θ, denoted by $\tilde{\theta}$, we obtain the linear model

$$Y - P(\tilde{\theta}) = \frac{\partial P(\theta)}{\partial \tilde{\theta}} (\theta - \tilde{\theta}) + [R(\tilde{\theta}) + e] \qquad (2.14)$$

where $\frac{\partial P(\theta)}{\partial \tilde{\theta}}$ denotes the $(r^2-1) \times r$ matrix of partial derivatives of $P(\theta)$ with respect to the r elements of θ, evaluated at $\tilde{\theta}$; and $R(\tilde{\theta})$ denotes the vector of remainder terms in the Taylor expansion of $P(\theta)$ about the value of $\tilde{\theta}$. Possible initial estimations for the elements of θ are

$$\tilde{P}_i = \sum_{j=1}^{r} (p_{ij}+p_{ji})/2, \quad i=1,2,\ldots,r-1 \qquad (2.15)$$

$$\tilde{\alpha} = \{ \sum_{j=1}^{r} (p_{jj}-\tilde{P}_j^2)/[\tilde{P}_j(1-\tilde{P}_j)r] \}^{1/2}. \qquad (2.16)$$

The estimator $\tilde{P}_i$ is unbiased for P_i under the assumptions of the unbiased response model. The initial estimator (2.16) for α is suggested because the quantities, $(P_{ii}-P_i^2)/P_i(1-P_i)$, for all $i=1,2,\ldots,r$, are equal to α^2 for the unbiased response model (2.1, 2.2).

We estimate the vector $\theta - \tilde{\theta} = \epsilon$ by

$$\hat{\epsilon} = (F' \tilde{V}^{-1} F)^{-1} F' \tilde{V}^{-1} W \qquad (2.17)$$

where $W = Y - P(\tilde{\theta})$; $F = \frac{\partial P(\theta)}{\partial \tilde{\theta}}$; and

$$\tilde{V} = \frac{1}{n} \{ \mathrm{Diag}[P(\tilde{\theta})] - P(\tilde{\theta})[P(\tilde{\theta})]' \} . \qquad (2.18)$$

We consider the improved estimator

$$\hat{\theta} = \tilde{\theta} + \hat{\epsilon} \qquad (2.19)$$

and estimate its covariance matrix by

$$\hat{\mathrm{Cov}}(\hat{\theta}) = (F' \tilde{V}^{-1} F)^{-1} . \qquad (2.20)$$

An approximate test for the hypotheses (2.1, 2.2) of the unbiased response model is obtained with the statistic

$$X_U^2 = \sum_{i=1}^{r} \frac{[n_{ii} - nP_{ii}(\hat{\theta})]^2}{nP_{ii}(\hat{\theta})} + \sum_{i<j}^{r} \sum^{r} \frac{[n_{ij}+n_{ji}) - 2nP_{ij}(\hat{\theta})]^2}{2nP_{ij}(\hat{\theta})} \qquad (2.21)$$

where the $P_{ii}(\hat{\theta})$ and $P_{ij}(\hat{\theta})$, $i \neq j$, denote the estimates for the expected proportions (2.10, 2.11) obtained with the parameter estimates of (2.19). It can be shown [see Battese (1973)] that the statistic, X_U^2, converges to a central chi-square random variable with $\frac{1}{2}(r-2)(r+1)$ degrees of freedom under the hypothesis of the unbiased response model (2.1, 2.2).

3. EMPIRICAL EXAMPLE

During September and October of 1970 the Statistical Laboratory of Iowa State University conducted a survey of 262 Iowa farm operators. Each farm operator was personally visited in September and asked questions about his farming operations. About one month later each farm operator was personally visited by another interviewer. The questionnaire used for the second interview was constructed so that some of the questions were exactly the same as in the first interview. One of the purposes of the survey was to estimate the relative magnitude of the variance of response errors for several variates important in farm surveys. An analysis of the survey is presented in Battese, Fuller and Hickman (1972).

One of the questions that was asked farm operators in this study was: " In terms of total value of sales, what was the most important agricultural product sold from the land you operated in 1969?" Not all farm operators gave the same answer in the two interviews. We consider the data obtained in coding the responses into three categories of " most important product" : hogs, cattle, and not hogs or cattle. The distribution of the survey responses in the two different interviews is shown in Table 1.

Table 1. Frequency of farmers reporting the " most important product"

	Trial-2 class			
Trial-1 class	Hogs	Cattle	Other	Totals
Hogs	85	9	2	96
Cattle	12	77	4	93
Other	9	8	56	73
Totals	106	94	62	262

With these sample frequencies, the statistic (2.9) to test for " lack of independent classifications in the two trials" has the value

$$X_S^2 = (9-12)^2/21 + (2-9)^2/11 + (4-8)^2/12 = 6.22 .$$

The five percent critical value for a chi-square distribution with three degrees of freedom [$\frac{1}{2}(r^2 - r) = 3$ for $r = 3$] is 7.81. At this level we do not reject the hypothesis of independent classifications in the two trials of the survey.

The initial estimators (2.15, 2.16) for the parameters in the unbiased model have values $\tilde{P}_1 = 0.385$, $\tilde{P}_2 = 0.357$ and $\tilde{\alpha} = 0.864$. From these initial estimates for the parameters in the model, the estimates for the P_{ij} in (2.10, 2.11) are $\tilde{P}_{11} = 0.325$, $\tilde{P}_{12} = 0.035$, $\tilde{P}_{13} = 0.025$, $\tilde{P}_{22} = 0.299$ and $\tilde{P}_{23} = 0.023$. The variables involved in the estimator $\hat{\epsilon}$ of (2.17) are thus

$$W = \begin{bmatrix} 0.324 \\ 0.034 \\ 0.008 \\ 0.046 \\ 0.294 \\ 0.015 \\ 0.034 \\ 0.030 \end{bmatrix} - \begin{bmatrix} 0.325 \\ 0.035 \\ 0.025 \\ 0.035 \\ 0.299 \\ 0.023 \\ 0.025 \\ 0.023 \end{bmatrix} = \begin{bmatrix} -0.001 \\ -0.001 \\ -0.018 \\ 0.011 \\ -0.005 \\ -0.008 \\ 0.009 \\ 0.007 \end{bmatrix}$$

and

$$F = \begin{bmatrix} 0.942 & 0.000 & 0.409 \\ 0.091 & 0.098 & -0.238 \\ -0.032 & -0.098 & -0.172 \\ 0.091 & 0.098 & -0.238 \\ 0.000 & 0.927 & 0.397 \\ -0.091 & -0.025 & -0.159 \\ -0.032 & -0.098 & -0.172 \\ -0.091 & -0.025 & -0.159 \end{bmatrix}.$$

The estimated covariance matrix (2.18) is obtained with the values of the vector $\hat{P}(\hat{\theta})$. From these data the elements of the estimator $\hat{\epsilon}$, defined by (2.17), are calculated to be 0.0003, 0.0030 and -0.0009 with standard errors 0.028, 0.027 and 0.020, respectively. The new estimates for the parameters in the model are thus

$$\hat{P}_1 = 0.386, \quad \hat{P}_2 = 0.360 \text{ and } \hat{\alpha} = 0.863.$$

With these parameter estimates the expected frequencies for the two interviews are estimated by $n\hat{P}_{11} = 85.25$, $n\hat{P}_{12} = 78.97$, $n\hat{P}_{13} = 53.92$, $n\hat{P}_{22} = 18.55$, $n\hat{P}_{23} = 13.10$, and $n\hat{P}_{33} = 12.21$. The statistic (2.21) for testing the response model has the value

$$\begin{aligned} X_U^2 &= (85-85.25)^2/85.25 + (77-78.97)^2/78.97 \\ &\quad + (56-53.92)^2/53.92 + (21-18.55)^2/18.55 \\ &\quad + (11-13.10)^2/13.10 + (12-12.21)^2/12.21 \\ &= 0.79. \end{aligned}$$

The five-percent critical value for the chi-square distribution with two degrees of freedom $[\frac{1}{2}(r-2)(r+1) = 2 \text{ for } r = 3]$ is 5.99. We therefore conclude that the unbiased response model (2.1, 2.2) is consistent with the observed frequencies for the "most important agricultural product in 1969."

4. EXISTENCE OF GENERAL UNBIASED RESPONSE MODELS

The unbiased response model, defined by (2.1, 2.2), consists of (r-1) independent population proportions and one response parameter, α. The model has the property that the probabilities of incorrectly reporting a given class are the same. We seek to determine if there exist more general response models that satisfy the unbiasedness conditions;

$$\sum_{m=1}^{r} P_m \beta_{mi} = P_i, \quad i = 1, 2, \ldots, r.$$

We assume that for a survey sample, in which each individual reports his classification in an interview and a re-interview, the expected proportions are (conceptually) known and satisfy the conditions

$$P_{ij} = P_{ji}, \quad i, j = 1, 2, \ldots, r \tag{4.1}$$

and

$$\sum_{j=1}^{r} P_{ij} = P_i, \quad i = 1, 2, \ldots, r. \tag{4.2}$$

We seek to determine conditions under which it is possible to "recover" the response probabilities that generated the expected proportions, P_{ij}.

We consider the equations (4.1, 4.2) and

$$P_{ij} = \sum_{m=1}^{r} P_m \beta_{mi} \beta_{mj}, \quad i, j = 1, 2, \ldots, r, \tag{4.3}$$

and seek to solve for the parameters β_{ij}, $i, j = 1, 2, \ldots, r$, such that they are nonnegative and satisfy the conditions $\sum_{j=1}^{r} \beta_{ij} = 1$, for all $i = 1, 2, \ldots, r$.

We first investigate the solution for the case of two classes (r=2). For the two-category case the conditions, $\sum_{j=1}^{2} \beta_{ij} = 1$ and $\sum_{m=1}^{2} P_m \beta_{mi} = P_i$, $i = 1, 2$, imply that the response parameters, β_{12}, β_{21} and β_{22} are expressed by

$$\beta_{12} = 1 - \beta_{11} \tag{4.4}$$

$$\beta_{21} = (1-\beta_{11})P_1/(1-P_1) \tag{4.5}$$

and

$$\beta_{22} = 1-(1-\beta_{11})P_1/(1-P_1) \tag{4.6}$$

The expected proportion P_{11}, expressed in terms of β_{11}, is thus

$$P_{11} = \beta_{11}^2 P_1 + (1-\beta_{11})^2 P_1^2/(1-P_1). \tag{4.7}$$

By expressing this equation as a quadratic in

β_{11} we obtain

$$0 = P_1(\beta_{11}-P_1)^2-(1-P_1)(P_{11}-P_1^2).$$

There exists a real solution for β_{11} if

$$P_{11}-P_1^2 \geq 0 . \qquad (4.8)$$

Given that this condition is satisfied a solution for the response parameters is

$$\beta_{ii} = |c| + (1-|c|)P_i , \quad i = 1, 2 \qquad (4.9)$$

$$\beta_{ij} = (1-|c|)P_j , \quad i \neq j \qquad (4.10)$$

where

$$c^2 = (P_{11}-P_1^2)/P_1(1-P_1)$$

$$= (P_{22}-P_2^2)/P_2(1-P_2) . \qquad (4.11)$$

Further, if the expected proportions satisfy the conditions

$$P_i^2 \leq P_{ii} \leq P_i^2/(1-P_i), \quad i = 1, 2, \qquad (4.12)$$

then the solutions for the response parameters are

$$\beta_{ii} = c + (1-c)P_i , \quad i = 1, 2 \qquad (4.13)$$

$$\beta_{ij} = (1-c)P_j , \quad i \neq j , \qquad (4.14)$$

where c^2 is defined by (4.11). It follows from (4.13, 4.14) that if the negative root of (4.11) is taken when the conditions of (4.12) are satisfied, then β_{ii} is less than β_{ji}, $j \neq i$, $i = 1, 2$. This is an unlikely situation in practice so that the solution defined by (4.9, 4.10) gives the appropriate response probabilities for an unbiased response model in the two-category case. The response model defined by (4.9, 4.10) is obviously a member of the class of unbiased response models defined by (2.1, 2.2).

It is readily seen that when there are more than two categories for the responses, solutions of the Equations (4.1, 4.2, 4.3) for the response parameters cannot be obtained in closed form without additional assumptions. We assume that for the r-category case ($r \geq 3$) Condition 1 is satisfied.

Condition 1: The probabilities of correct classification for class i, $i = 1, 2, \ldots, r$, are those that would be obtained from the 2 x 2 interview-reinterview problem considering only the two classes, "class i" and "not class i."

Given Condition 1 it can be shown [see Battese (1973)] that the solution for the probability of a correct response for class i is

$$\beta_{ii} = P_i + [(1-P_i)(P_{ii}-P_i^2)/P_i]^{1/2} \qquad (4.15)$$

provided that P_{ii} is no smaller than P_i^2. Further, if $r = 3$ and Condition 1 is satisfied, then the solution for the probability of incorrectly reporting class j when class i is the true class is

$$\beta_{ij} = (1-\beta_{jj})P_j/(1-P_j), \quad i \neq j . \qquad (4.16)$$

For the case when $r > 3$ we assume that Condition 2 is satisfied.

Condition 2: The probabilities of misclassification, β_{ij}, $i \neq j$, $i, j = 1, 2, \ldots, r$, are those that would be obtained from the 3 x 3 interview-reinterview problem considering only the three classes, "class i", "class j", and "not class i or class j."

Given Conditions 1 and 2 the solutions for the response probabilities are given by (4.15) and (4.16). These solutions and the conditions $\sum_{j=1}^{r} \beta_{ij} = 1$, for all $i = 1, 2, \ldots, r$, imply that the relationship

$$(P_{ii}-P_i^2)/P_i(1-P_i) = P_{jj}-P_j^2)/P_j(1-P_j) = c^2 \qquad (4.17)$$

holds for all i and j. Equations (4.15)-(4.17) imply that if the response probabilities satisfy Conditions 1 and 2 then the solution of the response probabilities in terms of the expected proportions is given by

$$\beta_{ii} = |c| + (1-|c|)P_i, \quad i = 1, 2, \ldots, r \qquad (4.18)$$

and

$$\beta_{ij} = (1-|c|)P_j, \quad i \neq j, \; i, j = 1, 2, \ldots, r \qquad (4.19)$$

where c^2 is defined by (4.17)

Conditions 1 and 2 are strong conditons and it is easy to think of situations where we would expect them to be violated. For example, we would not expect the situation to hold for individuals placed into classes on the basis of a free response to a continuous variable. However, in situations where sample responses are obtained to open-end questions, such as," What is the most important source of your income?", it may be reasonable to assume that the response hypotheses satisfy Conditions 1 and 2.

5. CONCLUSIONS

In this paper we present a class of unbiased response models that is defined by (r-1) independent proportions and a single response parameter. The response probabilities are a function of the true class proportions and the probabilities of misclassification in a given class are the same. Models that define the response probabilities independently of the population proportions generally will not satisfy the unbiasedness condition. For example, Mote and Anderson (1965), considered two simple response models in their investigation of the effect of misclassification on the size and power of chi-square goodness-of-fit tests for categorical data. The first model assumed equal probabilities of misclassification and the second model assumed that the only misclassifications were into classes adjoining the true classes. In both of these models the sample proportions are not, in general, unbiased estimators of the corresponding population proportions.

ACKNOWLEDGEMENT

The work reported in this paper was partially supported by the Bureau of the Census through the Joint Statistical Agreement, J.S.A. 72-4.

REFERENCES

[1] Assakul, K. and Proctor, C.H. (1967), "Testing Independence in Two-Way Contingency Tables With Data Subject to Misclassification," Psychometrika, 32, 67-76.

[2] Battese, G.E. (1973), "Parametric Models for Response Errors in Survey Sampling," Ph.D. thesis, 127 p., Iowa State University, Ames.

[3] Battese, G. E., Fuller, W. A. and Hickman, R. D. (1972), "Interviewer Effects and Response Errors in a Replicated Survey of Farm Operators in Selected Iowa Counties," Unpublished Report to the USDA's Statistical Reporting Service, Statistical Laboratory, Iowa State University, Ames.

[4] Bershad, M.A. (1967), "Gross Changes in the Presence of Response Errors," Unpublished memorandum, U.S. Bureau of the Census, Washington, D.C.

[5] Bross, I. (1954), "Misclassification in 2x2 Tables," Biometrics, 10, 478-86.

[6] Giesbrecht, F. G. (1967), "Classification Errors and Measures of Association in Contingency Tables," Proceedings of the Social Statistics Section of the American Statistical Association, 1967, 271-76.

[7] Huang, H. T. (1972), "Combining Multiple Responses in Sample Surveys," Ph.D. thesis, Iowa State University, Ames.

[8] Koch, G.G. (1968), "A Simple Model for Misclassification Errors in 2x2 Contingency Tables," Research Triangle Institute Technical Report SU-363, 24-28.

[9] Koch, G.G. (1969), "The Effect of Nonsampling Errors on Measures of Association in 2x2 Contingency Tables," Journal of the American Statistical Association, 64, 852-63.

[10] McCarthy, P.J. (1972), "Effects of Discarding Inliars When Binomial Data Are Subject to Classification Errors," Journal of the American Statistical Association, 67, 515-29.

[11] Mote, V.L. and Anderson, R.L. (1965), "An Investigation of the Effect of Misclassification on the Properties of χ^2-tests in the Analysis of Categorical Data," Biometrika, 52, 95-109.

THE MEASUREMENT OF FACTORS OF COMPREHENSION IN READING

Ann M. Bezdek, University of Illinois

I. Introduction

Reading comprehension has long been considered to be one of the most important goals of reading instruction. Yet relatively little is known of the nature of this complex ability even though much research has been devoted to the study of the nature of reading comprehension in the past three decades.

It has been theorized that comprehension in reading is composed of many underlying skills or abilities and that in order to comprehend what is read an individual must utilize numerous varied mental processes. Although reading theorists have generally described the process of reading comprehension as consisting of a number of specific skills, the results of studies concerned with identifying component skills of comprehension in reading to date have proved to be inconclusive. Many of these studies have utilized factor analytic techniques and while some studies have concluded that reading comprehension can be regarded as a unitary skill, others have provided evidence to indicate that some comprehension skills may be specific and unique, thereby supporting a multi-factor theory of reading comprehension.

Interestingly, the majority of studies designed to investigate the components of comprehension in reading have examined comprehension skills in subjects at the high school or college level where the reading process is likely to be a complex and highly organized process. The importance of examining the nature of the process at the elementary school level has been emphasized by Lennon (1962), Davis (1968), Bezdek (1973) and others. The present study was designed to investigate the specific factors of comprehension in reading to determine the unique measurable aspects of reading comprehension at the elementary level. This paper describes the manner in which the research was conducted and the data analyzed. Section 2 is concerned with a review and synthesis of the more notable studies that employed a factor analytic solution to examine the component abilities of reading comprehension. Section 3 describes the skills selected for measurement and the methods used to develop the items. A description of the test administration procedures, and an explanation of the manner in which the items were analyzed and the test revised is the content of section 4. Section 5 is concerned with the experimental test results and an analysis of the subtests' ability to function as measures of separate abilities. The results of the factor analysis are also presented in this section. Finally, the conclusions are found in section 6.

II. Earlier Factor Analytic Studies of Reading Comprehension

One of the earliest factor analytic studies concerned with comprehension in reading was reported by Feder (1938). The Comprehension Maturity Test designed to measure the ability to answer factual questions, the ability to make inferences from the material read, appreciation of the material read, and speed of reading was administered to 99 college sophomores. The common variance of the variables was factor analyzed using Thurstone's centroid method of factor analysis. The results were then rotated orthogonally. Feder's analysis of the results led him to conclude that reading for information and reading for inference were two distinct reading abilities.

Two years after Feder reported his findings Gans (1940) designed a study to determine the relationship between general and critical reading comprehension. Several published tests were used as a measure of general reading comprehension, and as a measure of critical reading skills, Gans constructed the Test of Reading Selection-Rejection. The tests were administered to a group of 417 pupils in grades four, five, and six in two New York Schools. The centroid method of factor analysis was applied to the results and rotated to oblique solutions. Gans concluded that reference reading is composed of general reading ability, and the ability to determine the relevance of designated sentences, and that it also involved some function of delayed recall.

Langsam (1941) using eight published reading tests obtained the intercorrelations of 21 variables in a group of 100 female college freshman. After the application of the centroid method of factor analysis and a rotation of the axes, Langsam found four factors to be specific to the skill of reading comprehension. The factors were identified as a verbal factor concerned with word knowledge, a perceptual factor involved with the speed of the response, a word factor described as fluency in dealing with words, and a factor tentatively identified as seeing relationships. A fifth factor identified as a number factor was not basic to the reading process.

Perhaps the most notable of the factor analytic studies concerned with reading comprehension was reported by Davis (1941). Davis reviewed the literature and identified nine skills considered to be the most important by reading theorists:

1. Knowledge of word meanings.
2. Ability to select the appropriate meaning for a word or phrase in the light of its particular contextual setting.
3. Ability to follow the organization of a passage and to identify antecedents and references in it.
4. Ability to select the main thought of a passage.
5. Ability to answer questions that are specifically answered in a passage.
6. Ability to answer questions answered in a passage but not in words in which the question is asked.
7. Ability to draw inferences.
8. Ability to recognize the literary devices used in a passage and to determine its tone and mood.
9. Ability to determine a writer's purpose, intent, and point of view.

Davis constructed items to measure each specific skill and assembled the 240 items into a test which was administered to 421 freshmen college students in several eastern colleges. A principal-axis component analysis was performed on the matrix of obtained score variances and covariances of the nine skill scores. Davis found six skills to be independent: word knowledge, the ability to reason in reading, the ability to follow the organization of a passage, the ability to recognize literary devices, the ability to understand a writer's explicit statements, and the ability to identify the writer's purpose. Davis explained that only word knowledge and the ability to reason in reading were measured with sufficient reliability to warrant their use for practical purposes.

Conant (1942) investigated the nature of reading comprehension to determine if the skill is best represented as a unitary ability or if it is composed of numerous skills. She constructed a test to measure six skills and administered it to 256 high school students and 74 freshmen college students along with the Nelson Denny Reading Test and the American Council Psychological Exam. Conant found high intercorrelations of the subtest scores and after applying Hotelling's principal components method of factor analysis she maintained that reading comprehension is a unitary skill.

Hunt (1952) reexamined the factors involved in reading comprehension originally identified in Davis' 1941 study. A pool of items from the Cooperative Reading Tests were classified by a group of consultants according to the verbal definition of each skill. The resulting 224 item test was administered to 560 college students. After the tests were scored Hunt determined the internal consistency of each skill measure, the intercorrelations among the six skill measures, and finally applied a factor solution to appraise whether the variance common to the relationhips among the skill measures could be accounted for by a general factor of reading comprehension. Hunt's results indicated that reading comprehension is comprised of a general reading factor and a word factor.

In another experiment, Singer (1962) reported a study that explored the factors of reading ability at the fourth grade level. He constructed measures of 36 variables and administered the tests to 60 fourth grade pupils. Employing the "substrata-factor analysis" Singer identified three factors of reading; word recognition, work meaning, and reasoning in context.

More than 25 years after his original study, Davis (1968) reexamined the uniqueness of the comprehension skills he identified in 1941. Selecting items from a pool of items that were not used in the published form of the Cooperative Reading Comprehension Tests, Davis developed two forms of a test to assess comprehension skills. Following a trail testing, the items underwent a differential item analysis and from the original 40 items measuring each skill, 24 were selected for use in the final test. Final forms of the test were administered to 988 senior high school students. The test scores were used to perform a uniqueness analysis cross validated by items and by examinees. Davis found that two factors, vocabulary knowledge and drawing inferences, were relatively unique comprehension skills. These findings were similar to Davis' 1941 findings where two skills, memory for word meanings and reasoning in reading, provided the largest factor loadings.

Schriener (1968) sought to determine the extent to which various comprehension subtests actually measure unique aspects of reading comprehension at the elementary school level. Schriener either constructed items or selected items from published tests to measure eight skills. The experimental battery of tests was split into halves and both halves were administered to 513 fifth grade students. Four subtests from the Iowa Test of Basic Skills were included in the analysis. The intercorrelations between all 12 variables were computed and utilized to obtain an estimate of the uniqueness of the subtests. The reliabilities of the differences between all possible pairs of subtests were analyzed and the principal components method of factor analysis was applied and followed by a varimax rotation. Schriener found that speed of reading, listening comprehension, verbal reasoning, and speed of noting details were relatively independent skills. From the factor solution however, speed of reading was the only identifiable factor involved in reading comprehension.

From a review of the research concerned with identifying the components of reading comprehension it is apparent that there exists no conclusive evidence to support either a one factor or a multi-factor theory of reading comprehension. In several instances experimentors felt that evidence was secured to indicate a high degree of specificity from one skill

measure to another when in some cases skills were highly correlated and overlapped.

Despite the varying interpretations of the research findings it appears that there is some evidence of several factors in reading comprehension, a word factor, a reasoning factor, and a literal understanding factor have been indicated by several studies. The great disparity in the conclusions reached in these studies can be attributed to the various instruments employed to assess the skills in question, the various statistical techniques applied, the different combinations of skills selected for analysis, differences in characteristics and ages of the subjects, and the manner in which the empirical results were interpreted.

III. Skill Selection and Item Construction

After careful analysis of the empirical studies that have been concerned with the analysis of reading comprehension, six of the eight skills analyzed by Davis (1968) were selected for investigation:

1) Remembering word meanings
2) Inferring word meanings from the context
3) Understanding content stated explicitly
4) Understanding the main idea of a passage
5) Making inferences about the content
6) Recognizing the author's purpose and point of view

The first skill, remembering word meanings, involves vocabulary knowledge or the ability to understand word meanings. The second skill is concerned with the ability to determine the meaning of unknown words by using the context for clues. Understanding the content stated explicitly, skill three, is concerned with literal comprehension and involves the student's ability to comprehend facts and details. The fourth skill pertains to the ability to understand the main idea of a selection. Skills five and six can be termed critical reading skills: skill five requires the ability to interpret what is read and draw conclusions based upon the facts presented, while skill six involves the ability to determine why an author uses the words he does, to recognize his bias and prejudices, and to interpret his feelings.

It should be pointed out that Davis in 1968 examined these skills at the high school level and before investigating these skills at the intermediate grade level some proof of their appropriateness for examination at this level was needed. A review of studies concerned with reading and thinking abilities of elementary school children was undertaken and this review indicates that the six skills adopted for investigation were appropriate skills for analysis in the intermediate grades.

Items to measure each of the comprehension skills were carefully constructed to measure the skills under investigation. Each item was based on a separate passage at the appropriate level of difficulty in order to eliminate any spurious interrelationships of item scores.

After the items were constructed they were submitted to a group of reading and measurement experts for analysis and were subsequently revised or eliminated on the basis of the suggestions offered by the team of analysts.

The revised items were then assigned to either Form A or Form B, the two trial test forms of the Reading Comprehension Skills Test. Each trail form consisted of six subtests of 14 items each designed to measure the six reading comprehension skills.

IV. Trail Testing and Item Analysis

In March of 1972, each trail form of the Reading Comprehension Skills Test was administered to a group of 50 fifth grade students on consecutive days. Students participating in the tryout testing were located in four fifth grade classrooms in a public elementary school in Champaign, Illinois and were selected because of their availability. Students were told the purpose for reading before each subtest and were provided with several sample items. Ample time was provided for students to complete all items, but several students did not do so.

While taking the test students were monitored and the protocols of students who were blindly marking answers were excluded from the item analysis along with the protocols of students who failed to answer each item. After excluding these protocols, the scores from 45 students who took Form A and 47 students who took Form B were used for the purpose of item analysis.

To identify items for inclusion on the final experimental form of the Reading Comprehension Skills Test correlation coefficients between pass or fail on each item in Form A and the total scores for each subtest in Form A were computed. Each item was excluded from the total score of the subtest of which it was a part when it was correlated with that subtest score. In this manner, no coefficient was spuriously increased by including an item in the total score with which it was correlated. The same procedure was followed with the items in Form B, the other tryout test form.

The resulting correlations were carefully reviewed and from the 28 items measuring each of the six comprehension skills, 14 were selected for use on the final experimental test. The items selected for inclusion on the final test had a higher average correlation with the total score on the skill they were designed to measure than with the total scores on the other skills. These correlations are given in table 1. In

general the mean differences are small, as would be expected, but, nevertheless, the items for each skill had subjective and empirical justification for their inclusion.

TABLE 1
Median Correlation Coefficients of Items with Total Subtest Scores

Skill	Median Correlation Coefficient of 14 items with total score of the skill measured by the items	Median Correlation Coefficient of 14 items with total scores of skills not measured by the items
1	.39	.29
2	.28	.27
3	.42	.33
4	.45	.37
5	.34	.31
6	.33	.31

To determine the appropriateness of the selected test items for fifth grade students, difficulty indices were calculated for each item. These indices are expressed as the proportion of students passing each item. The range of the difficulty indices and the median difficulty index for each skill test are shown in Table 2.

For efficiency in measurement items of appropriate difficulty are those with difficulty indices ranging from .30 to .70. Table 2 indicates that the median difficulty index for each skill measure falls within this range.

TABLE 2
Percentage of Students Marking Items Correctly in Forms A and B

Skill	Range for 14 items	Median Percentage
1	331-81	63
2	36-79	67
3	51-83	65
4	43-69	63
5	40-78	61
6	36-72	49

In sum, item analysis data were used to identify the best items for inclusion in the final test form. The items selected were constructed and initially judged to measure the skill they were intended to measure. The items correlated more highly with the skill they were intended to measure than with the other skills and were of appropriate difficulty for fifth grade students.

V. Experimental Test Administration and Results

The final experimental test, Form C, of the Reading Comprehension Skills Test consisted of 84 items arranged into subtests of 14 items each to measure the six comprehension skills being investigated. The sample for the final testing consisted of 369 fifth grade students enrolled in five public elementary schools in Champaign, Illinois. Previously administered standardized tests indicated that the students were slightly above average in general reading achievement. (Average reading achievement grade score was reported as 5.5 on the Scholastic Testing Service's Educational Development Test administered at the beginning of the school year.) The students comprising the sample for this study were selected mainly because of their availability and not because they were representative of fifth grade students throughout the state or country. Students participating in the final testing did not participate in the tryout testing.

Form C of the Reading Comprehension Skills Test was administered to the 369 fifth grade students in May of 1972. The test was administered in two one hour sessions on consecutive days and students were again provided with an explanation of the purpose for reading prior to taking each of the subtests. Sample items were also provided prior to subtest administration. Students were observed while taking the test and the protocols from a small number of students which were incomplete or incorrectly marked were excluded from the final analysis.

Following the scoring of the tests the first step in the analysis was to determine the reliability of the skill measures. The reliability estimates are based upon one administration of one test form, therefore, the reliability was determined through the use of a type of split-half method. The Kuder-Richardson formula 21 provided by Gronlund (1965 p.85) was used to obtain the coefficients. The resulting reliability coefficients are shown in Table 3 along with the standard error of measurement for each subtest and for the total test.

TABLE 3
Reliability* and Standard Error of Measurement for the Six Skills on the Reading Comprehension Skills Test Form C

Skill	r	S_m
1. Vocabulary	.56	1.77
2. Infering Word Meaning	.64	1.72
3. Reading for Details	.80	1.47
4. Reading for the Main Idea	.75	1.61
5. Making Inferences	.72	1.64
6. Determining Author Purpose and Point of View	.70	1.72
Total Test	.94	4.05

*The estimate of reliability are based on the Kuder-Richardson Formula 21.

In general, the reliability coefficients for the six skill measures given in Table 3 are not sufficiently high to warrant any diagnostic interpretations of the test scores. It should be pointed out that the estimates in Table 3 are likely to be underestimates of the subtests' reliability since the KR-21 formula assumes that all items have equal difficulty. As a whole the test functions as a very accurate measure of reading comprehension.

The second step in the analysis was to obtain the intercorrelations among the skill measures. If each subtest in fact assesses a specific skill the intercorrelations among the various skill measures would tend to be low. Relatively high coefficients of correlation, on the other hand, would indicate that the subtests are measuring the same process. The intercorrelations among the subtests are given in Table 4.

TABLE 4
Intercorrelation Coefficients of Total Scores on Subtests

	1	2	3	4	5	6
1						
2	.61					
3	.65	.70				
4	.66	.69	.75			
5	.64	.65	.70	.71		
6	.61	.61	.70	.70	.69	

The correlation matrix in Table 4 indicates coefficients above .60 and although what one calls "high" correlation is rather arbitrary, in this case it seems doubtful that the original hypothesis will be supported. To further investigate the uniqueness of the six skill measures the principal components method of factor analysis was applied to the matrix of intercorrelations given in Table 4. The results of the factor analysis are given in Table 5.

TABLE 5
Results of Factor Analysis Using Principal Components Method

Skill	Factor I Test Loading	h^2	r	r - h
1	.811	.658	.560	-.098
2	.833	.694	.640	-.054
3	.885	.783	.800	.017
4	.882	.778	.750	-.018
5	.859	.738	.720	-.018
6	.842	.709	.700	-.009

An examination of Table 5 indicates that comprehension in reading for the group tested is a unitary skill. Although not shown in the table, the results of the factor analysis indicates that nearly 73 percent of the total variance was accounted for by Factor I. The test loadings, or correlations of the skill scores with Factor I, are given in the second column of the table. It is apparent that all six skill measures are highly loaded on Factor I and due to the nature of the skill measures Factor I can tentatively be identified as general reading comprehension.

To examine the uniqueness of each skill measure in order to determine if there was any specific variance for each skill measure that was unaccounted for by the general factor, the reliability of each skill measure which indicates the maximum part of the total test variance that can be explained by that skill, given in column 4 of Table 5, was compared to its communality estimate, which reveals the percent of total variance accounted for by the general factor, given in column 3. The communality estimate when substracted from the reliability estimate for each skill measure indicates the amount of specific variance unaccounted for. These values given in column 5 of Table 5, show that all of the test variance for each skill, given the reliability estimates, was accounted for by the general factor. Thus none of the skills have been found to be unique.

VI. Conclusion

The results of this study indicated that for the group tested, students who obtained high scores on one skill measure obtained high scores on the others as well. Thus there was a consistency of pupil ranking on the subtests and no evidence was obtained to indicate that students vary in their ability to utilize the six comprehension skills. Performance on the six skill measures was found to be a function of one general factor, tentatively identified as "general reading comprehension," which accounted for 73 percent of the total test variance. These results cannot be generalized to other groups of fifth grade students. It is possible that if the same items were administered to other groups of students a difference in the ability to apply each skill would be found. In this case the items would serve as differential measures of the six skills.

Suppose that subsequent administrations of the experimental test indicate that the subtests are not functioning to measure specific skills. Would this then imply that reading comprehension is a general unitary factor or that the component skills of reading comprehension are not the skills that have been investigated or that the measurement instrument was too crude? Until evidence is accumulated that leaves little doubt that reading comprehension is a general ability and is not composed of specific skills the search for the nature and specific skills involved in the process of comprehension in reading will undoubtedly continue.

The findings of this investigation did not provide evidence to support the multi-factor theory of reading comprehension. Because the analysis did not indicate any specific or unique test variance for the six skill measures no justification was provided to permit the interpretation of student scores on the subtests as indices of particular strength or weakness in the various skills. Thus the Reading Comprehension Skills Test could not be appropriately used as a diagnostic instrument, however, the findings suggest that the Reading Comprehension Skills Test is well suited to assess general reading comprehension ability. The total test reliability estimate indicates that the test accurately assesses comprehension ability and the items require students to read for a variety of purposes. It is desirable that general reading comprehension tests include items that require students to read for different purposes because of the influence testing has upon teaching. When items that require one to read for a variety of purposes appear on reading tests it is probable that instruction will be devoted toward developing these reading competencies. The Reading Comprehension Skills Test would be a valuable tool for these purposes.

The author knows of no other studies prior to the present one that investigated the components of reading comprehension at the elementary school level through the use of a test composed of items based upon separate passages that were carefully designed and constructed to measure specific comprehension skills. Future investigations of this type should continue to examine comprehension skills at the elementary school level and explore the construct validity of separate measures of comprehension skills included on experimental tests.

REFERENCES

Bezdek, A. M. The Assessment of Fundamental Skills Involved in Reading Comprehension. Unpublished doctoral dissertation, University of Illinois, 1973.

Conant, M. M. The Construction of a Diagnostic Reading Test, New York: Teachers College Bureau of Publications, Columbia University, 1942.

Davis, F. B. Fundamental Factors of Comprehension in Reading. Unpublished doctoral dissertation, Harvard University, 1941.

Davis, F. B. Fundamental Factors of Comprehension in Reading. Psychometrika, 1944, 9, 185-97.

Davis, F. B. Research in Comprehension in Reading. Reading Research Quarterly, 1968, 3, 499-545.

Feder, D. D. Comprehension Maturity Tests--A New Technique in Mental Measurement. Journal of Educational Psychology, 29, 597-606.

Gronlund, N. E. Measurement and Evaluation in Teaching. Toronto: Macmillan Co., 1965.

Hunt, L. C. A Further Study of Certain Factors Associated with Reading Comprehension. Unpublished doctoral dissertation, Syracuse University, 1952.

Langsam, R. S. Factorial Analysis of Reading Ability. Journal of Experimental Education, 10, 57-63.

Lennon, R. T. What Can be measured? The Reading Teacher, 1962, 15, 326-37.

Schreiner, R. L. A Study of Interrelations Among Different Approaches to Measuring Reading Comprehension. Unpublished doctoral dissertation, University of Iowa, 1968.

A STATISTICAL ANALYSIS OF SCHOLASTIC APTITUDES AND ACADEMIC PERFORMANCE IN THE COLLEGE OF BUSINESS ADMINISTRATION: THE MEMPHIS STATE EXPERIENCE

Charles Branyan, Memphis State University
Robert Dean, Memphis State University

INTRODUCTION

In essence, this paper deals with the problem of academic performance, i.e., how well students might be expected to perform once they get to college. More precisely it investigates the relationship between a student's scholastic aptitude and his or her academic performance over a four year period in a college of business administration. It is also concerned with the role core courses play in predicting over-all academic performance in the college of business. That is, do students who do well in the core courses also perform well in the remainder of the courses in their business program.

PREVIOUS RESEARCH

Lavin (1965) identifies ability, personality traits and socio-economic status as the major predictors of academic performance. It would appear, however, that personality factors contribute only modestly to the prediction of academic performance. For example, study habits and attitudes, considered key "personality" variables by many researchers, have been found by Birney and Taylor to have only a .29 correlation with college grades. Another personality factor, student interests, was found by Chronbach (1949) to have only a .19 correlation with the grade averages of freshmen.

The importance of socio-economic status as a linear predictor of college academic performance is also open to debate. Friedhoff's (1955) research suggests that much of the association between college grades and socio-economic status is eliminated when ability is controlled. In addition, Boyce (1956) and Davis (1956) found an inverse rather than a direct relationship between socio-economic status and college grades. These latter studies have focused attention on the performance differences between public and private school students and the findings suggest that students who come from very wealthy families have more interest in propriety than achievement, hence do not strive for higher grades.

Ability has generally been measured by high school grades, intelligence tests or college level ability tests. Chronbach (op. cit.) found that college level ability tests correlated in the .50-.55 range with grade point averages. Henry (1950) found correlations as high as .70 between aptitude tests and college grades although .50 was more common. Swensen (1957) found that high school grades were the best single predictor of college grades while Astin (1969) has concluded that a combination of high school grades and ability tests is the best estimator of college academic performance.

Most of the research that has been conducted on the correlates of academic performance has relied on what Lavin (op. cit.) refers to as global measures of ability and performance, i.e. single over-all measures such as aptitude test scores and final grade point averages. A few researchers such as Travers (1949), Fisher (1955) Horst (1957), and Astin (1969) have used several dimensions of aptitude to predict over-all grade averages as well as grades in specific courses. However, the superiority of this type of differential analysis has been disputed by Chronbach (op. cit.) and Berdie (1955) who claim that multifactor tests of ability add little to the prediction of academic performance beyond what the general aptitude factor will predict. Perhaps more important, most studies concerning the prediction of academic performance have focused on college freshmen or college students in general while little if any attempt has been made to develop performance predictors for students taking a particular subject area or curriculum. Such predictors could lead to a more efficient allocation of human capital among professional and white collar occupations and at the same time make the educational system more responsive to the needs and aspirations of individual students.

STUDY APPROACH

The objective of this study is to determine the degree of association between student intellectual ability and academic performance in a college of business administration. Ability is measured by college level entrance examination scores and the grade point average (G.P.A.) for seven core or foundation courses that are required of all business administration students.[1] Academic performance is measured in terms of the G.P.A. for all courses taken by the college of business administration student and the G.P.A. for economics and business courses only.

In determining the sample size for this study, it was assumed that the mean final grade plus and minus three standard deviations would include all the grades and that the maximum and minimum grade point averages were 4.0 and 2.0 respectively. Based on these assumptions, the range in final G.P.A. was 2.1 points and the standard deviation of the population mean was estimated to be 0.35.

To provide a narrow margin for error, a 99 per cent confidence level was established with the sample mean not allowed to vary more than 0.1 points from the population mean. This desired accuracy is obtained with a mean plus and minus 2.58 standard deviations. Given these parameters, N = 82.

The data for this study was obtained from the Admissions Office of Memphis State University. A table of four digit random numbers was used to select individual students, with the

first digit in the random number representing a particular file drawer and the last three digits the specific record within the drawer. Each record selected included a final G.P.A. based on 132 hours of courses. G.P.A.'s had to be calculated, however, for business and economic subjects as well as the seven core courses. It should be noted that the final G.P.A. includes only the last grade earned by a student in a particular course while the G.P.A.'s for business and economic subjects and the core courses include all grades received for a particular course.

STUDY FINDINGS

To test the hypothesis that academic performance in the college of business administration is dependent upon intellectual ability, the following arguments were specified:

1. Aptitude test scores are a measure of a student's general intelligence and problem solving capacity; therefore, they should have a strong positive association with final grade point averages.

2. The core courses in the college of business administration measure the student's ability to master basic concepts and principles in business, therefore, the grade point averages in these courses should have a strong positive association with final grade point averages.

Two different statistical tests were designed to answer the first argument. The first test consisted of dividing the student sample into "high ability" and "low ability" groups and then analyzing the differences in the mean grade point averages for (1) all courses taken in the college of business administration, (2) business and economic courses only, and (3) the seven core courses.[2] The results are summarized in Table 1. (All tables are at the end of this paper.)

As expected, Table 1 indicates that the high ability group out-performed the low ability group for each set of courses. This finding lends support to the argument that general learning ability accounts for much of the difference in academic performance regardless of the nature of courses or curriculum (see Chronbach, op. cit.). Table 1 also reveals that the mean G.P.A.'s for each set of courses are statistically significant at the 99% confidence level (i.e. the chance of obtaining the above listed Z scores for two groups within the same population is less than 1 in a hundred).

The second statistical test consisted of regressing aptitude test scores on G.P.A. scores for the above-mentioned sets of courses. The results of this test were inconclusive. Although the correlation coefficients had the right sign they were not statistically significant at the 95% confidence level. The regression coefficients also had the right sign but were highly unreliable predictors of G.P.A.'s. For the most part, the standard errors of the regression coefficients were as large or larger than the coefficients themselves. Because of the disappointing results with correlation and regression, it is difficult to assess the merits of aptitude test scores as a predictor of academic performance in the college of business administration. Although it seems clear enough that students with higher aptitude test scores also make higher grades in the college of business, it is also apparent that aptitude test scores are not a good linear predictor of G.P.A.'s in business subjects. The problem, it would seem, is one of precision and reliability rather than direction. That is, we know that students with good learning ability will perform quite well in business courses but aptitude test scores do not measure this performance accurately.

To determine if the relationship between the grade point average for the core courses and the over-all grade point average for business administration courses is positive and statistically significant at both the 5% and 1% confidence levels, multiple correlation and regression analysis was performed using the following variables:

Dependent variables

Y_1 = Final G.P.A. for all courses taken by the student in the college of business

Y_2 = Final G.P.A. for all business and economics courses

Independent variables (core courses)

X_1 = G.P.A. for MANAGEMENT 1010, INTRODUCTION TO BUSINESS

X_2 = G.P.A. for ACCOUNTING 2010, FUNDAMENTALS OF ACCOUNTING I

X_3 = G.P.A. for ACCOUNTING 2020, FUNDAMENTALS OF ACCOUNTING II

X_4 = G.P.A. for ECONOMICS 2110, PRINCIPLES OF ECONOMICS I

X_5 = G.P.A. for ECONOMICS 2120, PRINCIPLES OF ECONOMICS II

X_6 = G.P.A. for MANAGEMENT 2711, BUSINESS STATISTICS I

X_7 = G.P.A. for MANAGEMENT 3711, BUSINESS STATISTICS II

The results of the analysis when Y_1 is regressed on X_1-X_7 are summarized in Table 2. As Table 2 indicates, the core courses explain 53 per cent of the variation in the final G.P.A. The F test indicates that the ratio of explained to unexplained variance far exceeded what might be expected due to chance. Not all of the independent variables, however, are reliable predictors of final G.P.A. Using the rule of thumb that the regression coefficient must be more than twice its standard error to be reliable as a predictor, we find that both the statistics courses, the second course in economics principles and the first course in accounting principles fail to meet this test. One possible

explanation for the poor showing of these variables is that they do not really provide a learning foundation for courses taken outside the college of business. This reasoning takes on added significance when one realizes these latter courses make up roughly 40-60 per cent of the total course load for the business student.

To test this hypothesis, the non-business courses were eliminated from the calculation of the final G.P.A. and Y_2 was regressed on variables X_1-X_7. The results are shown in Table 3.

A comparison of tables 2 and 3 indicates that the coefficient of multiple correlation (R) and its square (R^2) increase markedly when non-business courses are dropped from the regression run. This result tends to support the argument that the business core courses have little or no learning foundation for non-business courses. The high R (.8781) and R^2 (.7710) also butress the argument that one can predict a student's final grade point average by determining how well he did in the core courses.

The reliability of the individual core courses as predictors of over-all grade performance also improved. In the second run, only X_6, Business Statistics I, failed to meet the "rule of thumb" test. X_6 had the wrong sign and in a test for collinearity, it was found that X_7 and X_6 had a correlation of .749, which tends to explain the latter variable's weak showing.

It is interesting to note that the influence of each independent variable on final G.P.A. varied considerably. Principles of Accounting II, Principles of Business, and Business Statistics II account for 87 per cent of the explained variance in the final G.P.A. On the other hand, Principles of Economics I and II and Business Statistics I account for only 13% of the explained variance. When major subject areas are combined, however, a slightly different picture emerges. More precisely, accounting has the greatest influence on final G.P.A., followed in order by economics, business statistics, and business principles.[3]

CONCLUSION

The results of the regression analysis support the argument that students' final G.P.A. in business and economic courses has a positive and high degree of association with the G.P.A. for the seven core courses. The argument that a positive and high degree of association exists between final G.P.A. and aptitude tests was not fully supported although it is apparent that students with high aptitude scores, when compared to students with low scores, obtain significantly higher final G.P.A.'s. The findings also suggest that the ability to handle basic business courses will not insure a high degree of proficiency in non-business courses, i.e. there is little or no learning transfer between business and non-business subjects. Said a little differently, demonstrated proficiency in the core business courses is not a reliable measure of general learning ability and aptitude; rather it is a measure of the student's ability and capacity to handle business and economics subjects only.

FOOTNOTES

[1]The seven core courses are as follows: Management 1010, Introduction to Business, Accounting 2010, Fundamentals of Accounting I, Accounting 2020, Fundamentals of Accounting II, Economics 2110, Principles of Economics I, Economics 2120, Principles of Economics II, Management 2711, Business Statistics I, Management 3711, Business Statistics II.

[2]High and low ability rankings were based on the mean aptitude test score for the above student sample. Those students with scores above the mean were placed in with the high ability group and those with scores below the mean were assigned to the low ability group. It should be noted that the mean score for the sample closely approximates the mean score for all freshmen students during the period under investigation.

[3]Due to a lack of space the regression run with the combined variables was not included in this paper. For further details see R. D. Dean and C. Branyan "Correlates of Academic Performance in the College of Business Administration." Working Paper No. 26, Memphis State University, Memphis, Tennessee 1973.

REFERENCES

[1] Astin, Alexander. Predicting Academic Performance in College, New York: The Free Press, 1971.

[2] Berdie, Ralph F. "Aptitude, Achievement, Interest and Personality Tests: A Longitudinal Study," Journal of Applied Psychology, Vol. 50, 1959.

[3] Birney, Robert C., and Marc J. Taylor. "Scholastic Behavior and Orientation to College," Journal of Educational Psychology, Vol. 50, 1959.

[4] Boyce, E. M. "A Comparative Study of Overachieving and Underachieving College Students on Factors other than Scholastic Aptitude," Dissertation Abstracts, Vol. 16, 1956.

[5] Chronbach, Lee J. Essentials of Psychological Testing, New York: Harper and Brothers, 1949.

[6] Davis, Junius A. "Differential College Achievement of Public and Private School Graduates," Journal of Counseling Psychology, Vol. 3, 1956.

[7] Fisher, Joseph T. "The Value of Tests and Records in the Prediction of College Achievement," Dissertation Abstracts, Vol. 15, 1955.

[8] Friedhoff, W. H. "Relationships Among

Various Measures of Socio-Economic Status, Social Class Identification Intelligence and School Achievement," Dissertation Abstracts, Vol. 15, 1955.

[9] Henry, Erwin R. "Predicting Success in College and University," in Fryer, Douglas H. and Erwin Henry, eds., Handbook of Applied Psychology, New York: Rinehart and Co., 1950.

[10] Horst, Paul. "Differential Prediction in College Admissions," College Board Review, Vol. 33, 1957.

[11] Lavin, David E. The Prediction of Academic Performance, New York: Russell Sage Foundation, 1965.

[12] Swensen, Clifford H. "College Performance of Students with High and Low High School Grades when Academic Attitude is Controlled," Journal of Educational Research, Vol. 50, 1957.

[13] Travers, Robert M. W. "Significant Research on the Prediction of Academic Research" in Donahue, W. T., et. al., ed., The Measurement of Student Adjustment Achievement, Ann Arbor, University of Michigan Press, 1949.

Table 1. Mean G.P.A.'s and Z Scores for High and Low Ability Groups

	High Ability Group Mean G.P.A. (X_1)	Low Ability Group Mean G.P.A. (X_2)	Z^I Scores	P
All Courses	2.54418	2.31941	2.8695*	.0041
Business/Economic Courses	2.58918	2.30862	2.7872*	.0053
Core Courses	2.50941	2.13751	3.0256*	.0024
Sample Size	38	43	*Significant at 99% level	

$$Z^I = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 - X_2}}$$

Table 2. Step-Wise Regression Analysis: Final G.P.A. for All Business and Non-Business Courses and Core Courses

Variable	Regression Coefficient	Multiple R	Multiple R^2	Increase R^2	F Ratio	Significant Points of F 5%	Significant Points of F 1%
X_3	.10044 (.04004)	.5126	.2627	.2627	28.508	3.96	6.96
X_1	.10559 (.04330)	.6478	.4196	.1569	28.561	3.11	4.88
X_7	.08724 (.04835)	.6845	.4685	.0489	22.919	2.72	4.04
X_4	.08716 (.04194)	.7121	.5071	.0386	19.804	2.48	3.56
X_2	.06474 (.03898)	.7228	.5225	.0154	16.631	2.33	3.25
X_5	.04066 (.04067)	.7271	.5287	.0062	14.020	2.22	3.06
X_6	-.02199 (.04857)	.7280	.5300	.0013	11.919	2.13	2.89
Constant	1.36106		N=82				

Table 3. Step-Wise Regression Analysis: Final G.P.A., Business and Economic Courses and Core Courses

Variable	Regression Coefficient	Multiple R	Multiple R^2	Increase R^2	F Ratio	Significant Points of F 5%	Significant Points of F 1%
X_3	.14417 (.03578)	.6257	.3915	.3915	51.469	3.96	6.96
X_1	.13331 (.03869)	.7657	.5898	.1983	56.790	3.11	4.88
X_7	.13048 (.03918)	.8215	.6748	.0850	53.954	2.72	4.04
X_5	.11833 (.03634)	.8477	.7186	.0438	49.163	2.48	3.56
X_2	.11216 (.03483)	.8682	.7538	.0352	46.532	2.33	3.25
X_4	.08734 (.03749)	.8780	.7708	.0171	42.084	2.22	3.06
X_6	-.00952 (.13048)	.8781	.7710	.0001	35.591	2.13	2.89
Constant	.78830		N=82				

SOME RATIO-TYPE ESTIMATORS

R. P. Chakrabarty, University of Georgia

1. INTRODUCTION

One of the main objectives of a sample survey is the estimation of the population mean or total of a characteristic 'y' attached to the units in the population. Ratio estimators are among the most commonly used estimators of the population mean or total of 'y' utilizing an auxiliary characteristic 'x' that is positively correlated with 'y'. The precision of the regression estimator is usually higher than that of the ratio estimator but in large-scale sample surveys, the ratio estimator is frequently employed because of its simplicity. In this paper, we develop some ratio-type estimators which will be more efficient than the customary ratio estimator and/or the unbiased estimator and yet computationally comparable to the customary ratio estimator.

We shall, without loss of generality, confine ourselves to the estimation of $\bar{Y}$, the population mean of 'y'. Further, to simplify the discussion, we shall confine ourselves to simple random sampling and assume the population size is infinite. From a simple random sample of n pairs (y_i,x_i) we have the unbiased estimator of $\bar{Y}$, as

$$\bar{y} = \sum_{i=1}^{n} y_i/n. \qquad (1.1)$$

The customary ratio estimator of $\bar{Y}$ is

$$\bar{y}_r = (\bar{y}/\bar{x})\bar{X} = r\bar{X} \qquad (1.2)$$

where $\bar{x}$ is the sample mean and $\bar{X}$ is the known population mean of x, and

$$r = \bar{y}/\bar{x} \qquad (1.3)$$

is the ratio estimator of the ratio $R = \bar{Y}/\bar{X}$.

It is well known that the ratio estimator $\bar{y}_r$ is more efficient than the unbiased estimator $\bar{y}$ in large samples if $\rho > C_x/(2C_y)$ where ρ is the coefficient of correlation between y and x and C_y and C_x are coefficients of variation of y and x respectively. The question of choice between $\bar{y}$ and $\bar{y}_r$ arises when it is suspected that $\rho(\geq 0)$ is not high and/or $C_x > C_y$. The customary procedure in such situations is to use $\bar{y}_r$ when $\rho > C_x/(2C_y)$ otherwise use $\bar{y}$. It is, however, desirable to develop alternative ratio-type estimators which are more efficient than $\bar{y}_r$ as well as $\bar{y}$ and yet computationally comparable to $\bar{y}_r$. The two ratio-type estimators we propose are

$$t_1 = (1-W)\bar{y} + W\bar{y}_r;\ W \geq 0 \qquad (1.4)$$

and

$$t_2 = (1-W)\bar{y} + W\, r^*\bar{x};\ W \geq 0 \qquad (1.5)$$

where W is a constant weight to be determined and

$$r^* = 2r - \tfrac{1}{2}(r_1+r_2) \qquad (1.6)$$

is obtained by splitting the sample at random into two groups, each of size n/2 and $r_j=\bar{y}_j/\bar{x}_j$, (j=1,2), $\bar{y}_j$ and $\bar{x}_j$ are means of y and x respectively obtained from jth half-sample. The estimator t_1 reduces to $\bar{y}$ and $\bar{y}_r$ when W=0 and 1 respectively. The estimator t_2 reduces to $\bar{y}$ when W=0 and when W=1 it reduces to $r^*\bar{X}$ which is the 'Jack-knife' ratio estimator of $\bar{Y}$. It may be mentioned here that by dividing the sample at random into g(≤n) groups, each of size n/g, a more general form of the estimator t_2 could be obtained as

$$t_{2g} = (1-W)\bar{y} + W[gr - \frac{g-1}{g}\sum_{j=1}^{g} r_j']\bar{X}$$

where r_j' is the customary ratio estimator calculated from the sample after omitting the jth group. However, in this paper we shall consider the special case of t_2 given in (1.5). Srivastava (1967) proposed the estimator

$$t_3 = \bar{y}(\bar{X}/\bar{x})^W \qquad (1.7)$$

where W is a constant weight and obtained its asymptotic variance. The estimator t_2 was suggested earlier by Chakrabarty (1968). In this paper these estimators will be compared regarding the properties of bias and efficiency. In section 2, we discuss the asymptotic theory and in section 3 we give the exact biases and variances of these estimators under a regression model.

2. ASYMPTOTIC THEORY

2.1 Biases of the estimators.

It is obvious that the estimators t_1, t_2, and t_3 are consistent but in general biased, like the ratio estimator $\bar{y}_r$. Now, as it is customary in the asympototic theory of ratio method of estimation, we shall assume that the sample size n is sufficiently large so that

$$|\delta_{\bar{x}}| = |\frac{\bar{x}-\bar{X}}{\bar{X}}| << 1 \qquad (2.1)$$

Under the above assumption, the expected value of r is given by

$$E(r) = R + \frac{R}{n}(C_x^2 - \rho C_y C_x) + O(n^{-2})$$

Now, since r_1 and r_2 are independent

$$E(r^*) = R + O(n^{-2})$$

Consequently, the biases of t_1 and t_2 are

$$\text{Bias}(t_1) = W\ \text{Bias}(\bar{y}_r) = \frac{W\bar{Y}}{n}(C_x^2 - \rho C_y C_x) + O(n^{-2}) \qquad (2.2)$$

and

$$\text{Bias}(t_2) = 0 + O(n^{-2})$$

respectively. From Srivastava (1967), the bias t_3 is given by

$$\text{Bias}(t_3) = \frac{W\bar{Y}}{n}[\frac{(W+1)}{2}C_x^2 - \rho C_y C_x] + O(n^{-2}) \qquad (2.4)$$

Thus, the asymptotic bias of t_2 is of order n^{-2} and hence smaller than that of $\bar{y}_r$, t_1 and t_3

whose biases are order n^{-1}. The bias of t_1 is smaller than that of $\bar{y}_r$ for $0<W<1$. We note that $C_x^2-\rho C_y C_x=0$ when the regression of y on x passes through the origin. Consequently, for the important case of regression through the origin the estimators $\bar{y}_r$ and t_1 are unbiased to terms of order n^{-1} but the bias of t_3 is still of order n^{-1}. Further, substituting the formula for exact bias of $\bar{y}_r$ from Hartley and Ross (1954) we get the exact bias of t_1 as

$$\text{Bias}(t_1) = -W\,\text{Cov}(r,\bar{x})$$

and

$$\frac{|\text{Bias}(t_1)|}{\sigma_t} \leq \frac{WC_x}{\sqrt{n}} \qquad (2.5)$$

Thus if $\frac{WC_x}{\sqrt{n}} \leq 0.1$, the bias of t_1 is negligible in relation to its standard error. No such upper bound to the bias of t_3 relative to its standard could be obtained.

2.2 Variances of the estimators.

In deriving the variances of estimators t_1, t_2 and t_3 we consider up to terms of n^{-1} only and biases which are of order n^{-1} are neglected. Expanding r and r_j by Tylor's series in terms of $\delta_{\bar{x}},\delta_{\bar{y}}$ and $\delta_{\bar{x}_j},\delta_{\bar{y}_j}$ $(j=1,2)$ it can be shown that to terms of order n^{-1} the variances of t_1, t_2 and t_3 are identical and are given by

$$V(t_1) = V(t_2) = V(t_3) = \frac{S_y^2}{n}[1+WK(WK-2\rho)] \qquad (2.6)$$

where $K = C_x/C_y$. (2.7)

The value of W which minimizes this variance is

$$W_{opt} = \rho/K \qquad (2.8)$$

The minimum variance is given by

$$V_{min} = \frac{S_y^2}{n}(1-\rho^2) \qquad (2.9)$$

which is equal to the variance of the linear regression estimator up to terms of order n^{-1}. Substituting W=1 in (2.6) we get the variance of $\bar{y}_r$ as

$$V(\bar{y}_r) = \frac{S_y^2}{n}[1+K(K-2\rho)] \qquad (2.10)$$

The asymptotic efficiencies of t_1 (t_2 and t_3) over $\bar{y}$ and $\bar{y}_r$ are given by

$$E_1 = \frac{V(\bar{y})}{V(t_1)} = \frac{1}{[1+WK(WK-2\rho)]} \qquad (2.11)$$

and

$$E_2 = \frac{V(\bar{y}_r)}{V(t_1)} = \frac{[1+K(K-2\rho)]}{[1+WK(WK-2\rho)]} \qquad (2.12)$$

respectively. From (2.11) and (2.12) we get

$$E_1 \geq 1 \quad \text{if} \quad W \leq 2\rho/K$$

and

$$E_2 \geq 1 \quad \text{if} \quad (2\rho-K)/K \leq W \leq 1 \qquad (2.13)$$

Thus the estimators t_1, t_2 and t_3 are better than $\bar{y}$ and $\bar{y}_r$ for a wide range of W-values. For example, if $\rho=.6$, $K=1$ and W is between 0.2 and 1 estimators t_1, t_2 and t_3 are asymptotically more efficient than $\bar{y}$ and $\bar{y}_r$. The efficiencies E_1 & E_2 of the estimators t_1, t_2 and t_3 over $\bar{y}$ and $\bar{y}_r$ will depend on ρ, K and the weight W. The numerical values of E_1 and E_2 for different values of ρ, K and for W=1/4 and W=1/2 are given as percentages in Tables 1 and 2 respectively. Comparing the results in the two tables we may conclude that if a good guess of ρ/K is not available from a pilot sample survey, past data or experience (1) W=1/4 appears to be a good overall choice for t_1, t_2 and t_3 for low correlation ($.2<\rho<.4$) and/or $K>1$. (2) W=1/2 appears to be a good choice for moderate to high correlation ($\rho>.4$) and $K>1$. (3) In cases where $\rho>.8$ and $K\leq 1$ it is preferable to use $\bar{y}_r$.

The asymptotic variance given in (2.9) of the estimators t_1, t_2 and t_3 with optimum value of $W=\rho/K$ is equal to the asymptotic variance of the linear regression estimator

$$\bar{y}_{\ell r} = \bar{y} + b(\bar{X} - \bar{x}) \qquad (2.14)$$

where b is the sample regression coefficient. Thus these estimators with constant weights (W=1/4 or 1/2) are asymptotically no more efficient than $\bar{y}_{\ell r}$. However, if the regression of y on x is not linear, Cochran (1963) has shown that the bias in $\bar{y}_{\ell r}$ is of order n^{-1} and hence it is more biased than t_2 whose bias is of order n^{-2}. Thus t_2 may be preferable to $\bar{y}_{\ell r}$ in situations where freedom from bias is important. Moreover, computationally t_2 is simpler than $\bar{y}_{\ell r}$.

3. THE EXACT THEORY

We assume the following model for the comparison of estimators:

$$y_i = \alpha + \beta x_i + u_i;\ \beta > 0$$

$$E(u_i|x_i) = 0,\ E[u_i, u_j|x_i,x_j) = 0$$

$$V(u_i|x_i) = n\delta \quad (\delta \text{ is a constant of order } n^{-1}) \qquad (I)$$

where the variates x_i/n have the gamma distribution with parameter h so that $\bar{x}=\sum x_i/n$ has the gamma distribution with the parameter $m=nh$. This model was used by Durbin (1959), and Rao and Webster (1966) to investigate the bias in estimation of ratios, and Chakrabarty and Rao (1967) to investigate the stability of the 'Jack-Knife' variance estimor in ratio estimation. Chakrabarty (1973) has used this model to investigate the exact efficiency of the ratio estimator $\bar{y}_r$ and

stability of the variance estimator of $\bar{y}_r$ relative to that of $\bar{y}$. He has shown that for $\rho \geq .4$ and $K < 2\rho$ the ratio estimator is generally more efficient than the unbiased estimator $\bar{y}$ even in small samples, and that the variance estimator of the ratio estimator is generally more stable than the variance estimator of $\bar{y}$. It may be noted that all our results under this model are exact for any sample size, n.

3.1 The exact biases of the estimators.
In terms of the model (I) we have

$$\bar{y} = \alpha + \beta\bar{x} + \bar{u}$$

$$E(\bar{y}) = \alpha + \beta m = \bar{Y}$$

$$t_1 = \alpha\left(1-W+\frac{Wm}{\bar{x}}\right) + \beta[(1-W)\bar{x} + Wm] + \bar{u}\left\{(1-W) + \frac{Wm}{\bar{x}}\right\} \qquad (3.1)$$

Consequently, the bias of t_1 is

$$\text{Bias}(t_1) = E(t_1) - (\alpha + \beta m) = \alpha W/(m-1) \qquad (3.2)$$

$$t_2 = \alpha\left[(1-W)+Wm\left(\frac{2}{\bar{x}} - \frac{1}{2\bar{x}_1} - \frac{1}{2\bar{x}_2}\right)\right] + \beta[(1-W)\bar{x}+Wm] - \frac{Wm}{2}\left(\frac{\bar{u}_1}{\bar{x}_1} + \frac{\bar{u}_2}{\bar{x}_2}\right) + \bar{u}\left[(1-W) + \frac{2Wm}{\bar{x}}\right]$$

$$E(t_2) = \beta m + \alpha[1-2W/(m-1)(m-2)]$$

Thus the bias of t_2 is

$$\text{Bias}(t_2) = -2W\alpha/(m-1)(m-2) \qquad (3.3)$$

$$t_3 = (\alpha+\beta\bar{x}+\bar{u})m^{W}\bar{x}^{-W}$$

$$E(t_3) = \frac{m^W}{\Gamma(m)}\left[\alpha\Gamma(m-W)+\beta\Gamma(m-W+1)\right]$$

Consequently, the bias of t_3 is

$$\text{Bias}(t_3) = \alpha\left[\frac{m^W\Gamma(m-W)}{\Gamma(m)} - 1\right] + \beta\left[\frac{m^W\Gamma(m-W+1)}{\Gamma(m)} - m\right] \qquad (3.4)$$

Now, putting W=1 in either (3.2) or (3.4) we get the bias of $\bar{y}_r$ as

$$\text{Bias}(\bar{y}_r) = \alpha/(m-1) \qquad (3.5)$$

From (3.2) through (3.5) it can be seen that the bias of t_2 is of order n^{-2} while those of $\bar{y}_r$, t_1 and t_3 are of order n^{-1} since m=nh in our model. Also, the bias of t_1 is less than the bias of $\bar{y}_r$ if W<1. Further, for the special case of the linear regression through the origin (i.e. α=0 in model I.) the estimators $\bar{y}_r$, t_1 and t_2 are unbiased but t_3 is still biased. A numerical evaluation of the biases of these estimators is given in the next section.

3.2 The exact variances of the estimators.
The method of obtaining exact expressions for the variances of these estimators under model I is similar to that of Rao and Webster (1966). The details of evaluating these variances, which involve some algebra, are omitted and only the final results are given here. The variance of t_1 can be shown to be

$$V(t_1) = \frac{W^2m^2}{(m-1)^2(m-2)}\alpha^2 + (1-W)^2m\beta^2 + \left[\frac{W^2m^2}{(m-1)(m-2)} + \frac{W(1-W)(m+1)}{(m-1)} + (1-W)\right]\delta - \frac{2W(1-W)m}{(m-1)}\alpha\beta \qquad (3.6)$$

Putting W=1 and W=0 in (3.6) the variance of $\bar{y}_r$ and $\bar{y}$ are obtained as:

$$V(\bar{y}_r) = \frac{m^2\alpha^2}{(m-1)^2(m-2)} + \frac{m^2\delta}{(m-1)(m-2)} \qquad (3.7)$$

and

$$V(\bar{y}) = \delta + \beta^2 m \qquad (3.8)$$

respectively. The variance of t_2 is obtained as

$$V(t_2) = \frac{W^2m^2(m^2-6m+17)}{(m-1)^2(m-2)^2(m-4)}\alpha^2 - \frac{2W(1-W)m(m-3)}{(m-1)(m-2)}\alpha\beta + (1-W)^2m\beta^2 + \left[(1-W)^2 + \frac{W^2(m^2-7m+18)m^2}{(m-1)(m-2)^2(m-4)} + \frac{2W(1-W)m(m-3)}{(m-1)(m-2)}\right]\delta \qquad (3.9)$$

Finally, the variance of t_3 is given by

$$[m^{-2W}\Gamma^2(m)]V(t_3) = [\Gamma(m-2W)\Gamma(m)-\Gamma^2(m-W)]\alpha^2 + [\Gamma(m+2-2W)\Gamma(m)-\Gamma^2(m+1-W)]\beta^2 + 2[\Gamma(m+1-2W)\Gamma(m)-\Gamma(m+1-W)\Gamma(m-W)]\alpha\beta + [\Gamma(m-2W)\Gamma(m)]\delta \qquad (3.10)$$

We note that in terms of the model I

$$\alpha = \bar{Y}[(K-\rho)/K]$$

$$\beta = \bar{Y}[\rho/(Km)]$$

$$\delta = \bar{Y}^2[(1-\rho^2)/(K^2m)] \qquad (3.11)$$

and $K = C_x/C_y$

The exact efficiencies of $\bar{y}_r$ and t_i (i=1,2, and 3), relative to that of $\bar{y}$ are given by

$$E_r' = V(\bar{y})/MSE(\bar{y}_r)$$

$$E_i' = V(\bar{y})/MSE(t_i) \qquad i=1,2 \text{ \& } 3 \qquad (3.12)$$

Now, using (3.2) through (3.10) and substituting the values of α, β and δ given by (3.11) efficiencies E_r' and E_i' (i=1,2&3) can be expressed explicitly as functions of $K=C_x/C_y$, m=nh, ρ and weight W. However, it is difficult to investigate analytically the efficiencies of the estimators from the resulting expressions. Therefore, we have evaluated the values of E_r' and E_i' (percentages) for selected values of ρ, K and m and for W=1/4 and 1/2. The results are given in Tables 3 and 4 respectively. The results of Table 3 may be sum-

marized as follows: (1) The ratio estimator $\bar{y}_r$ is less efficient than $\bar{y}$ for low correlation ($\rho \leq .4$) except when $\rho=.4$, $K<1$ and $m \geq 20$. (2) The estimators t_1, t_2 & t_3 with $W=1/4$ are more efficient than both $\bar{y}$ and $\bar{y}_r$ for the following values of ρ, K and m, (a) $.2 \leq \rho \leq .4$, $K \leq 1$, $m \geq 16$. (b) $.2 < \rho < .4$, $K>1$, $m \geq 32$. Noting that in our model $C_x = h^{-1/2}$ $C_x = m^{-1/2}$ and $n \leq m$ if $h \geq 1$ we may conclude that for low correlation ($.2 \leq \rho \leq .4$), $W=1/4$ appears to be a good choice for estimators t_1, t_2, & t_3 even in small samples if $K \leq 1$ and in large samples only when $K>1$. Further, the exact efficiencies of these estimators with $W=1/4$ are of the same order as judged by their mean square errors.

From table 4, it can be seen that the estimators t_1, t_2 and t_3 with $W=1/2$ are more efficient than both $\bar{y}$ and $\bar{y}_r$ for $\rho \geq .5$, $.25 \leq K \leq 1.50$ and $m \geq 16$. However, the ratio estimator $\bar{y}_r$ is most efficient when $\rho=.9$ and $.5 \leq K \leq 1$. Thus, $W=1/2$ appears to be a good choice for estimators t_1, t_2, and t_3 for moderate to high correlation ($\rho>.4$), except when $\rho=.9$ and $.5 \leq K \leq 1$. The exact efficiencies of t_1, t_2 and t_3 with $W=1/2$ are again generally of the same order. It is interesting to note that under model I the exact efficiencies of the estimators t_1, t_2 and t_3 approach the asymptotic efficiency when $m=nh \geq 32$. For example when $\rho=.4$ & $K=1.0$, $E_1=116$ (table 1) & $E_1'=114$, $E_2'=E_3'=115$ for $m=32$ (table 3).

We note from tables 3 and 4 that it is difficult to choose among the estimators t_1, t_2 and t_3 on the basis of their exact mean square errors. The absolute biases of estimators $\bar{y}_r$ and t_i relative to their mean square errors are given by

$$B_r = |\text{Bias}(\bar{y}_r)|/[\text{MSE}(\bar{y}_r)]^{1/2}$$

and

$$B_i = |\text{Bias}(t_i)|/[\text{MSE}(t_i)]^{1/2}, \quad i=1,2\&3 \qquad (3.13)$$

respectively. The numerical values of B_r and B_i ($i=1,2\&3$) for $W=1/4$ and $W=1/2$ are given in tables 5 and 6 respectively for selected values of m, K & ρ. From table 5, it can be seen that B_2 is generally less than 1%; B_1 is slightly greater than B_3 but B_1 is still less than 10% for $m=nh \geq 16$. The ratio estimator $\bar{y}_r$ is generally badly biased ($B_r>10\%$ for $K \geq 1$). From table 6, we find that B_2 $<1\%$ for $K \leq 1$ and for $K>1$, $B_2<2.5\%$ when $m \geq 16$. Turning to the relative biases of t_1 and t_3 we find that $B_1<B_3$ for $K<1$ and $B_1>B_3$ for $K>1$. It is also interesting to note that although $\text{MSE}(\bar{y}_r)<\text{MSE}(t_i)$ for $\rho=.9$ and $.5 \leq K \leq 1$ (table 4), B_r in this case exceeds 10% and is considerably higher than B_i. Thus, for $\rho=.9$ and $.5 \leq K \leq 1$, although $\text{MSE}(\bar{y}_r) < \text{MSE}(t_i)$, the estimators t_i's may be preferable in situations where the freedom from bias is desirable.

It may be noted that in surveys with many strata and small samples within strata the bias of the ratio estimator relative to its standard error may be considerable if it is appropriate to use 'separate' ratio estimators (see Cochran). In such situations it may be of great advantage to use the proposed estimators t_i ($i=1,2$ and 3). These estimators not only reduce the bias but also increase the precision.

In light of the above results we conclude that the three ratio-type estimators t_1, t_2 and t_3 are preferable to both $\bar{y}$ and $\bar{y}_r$. The efficiencies of these estimators are the same in large samples and are practically of the same order in small samples. Computationally t_1 is simplest and the bias of t_2 is least.

The author wishes to thank Dr. J. N. K. Rao for his valuable suggestions.

REFERENCES

Chakrabarty, R. P. and Rao, J. N. K. (1967). The Bias and Stability of Jack-Knife Variance Estimator in Ratio Estimation. Proc. Amer. Stat. Assoc. (Social Statistics Section).

Chakrabarty, R. P. (1968) Contributions to the Theory of Ratio-Type Estimators. Ph.D. Thesis, Texas A&M Univ.

Chakrabarty, R. P. (1973). A note on the Small Sample Theory of the Ratio Estimator in Certain Specified Populations, accepted for publication. Jour. Ind. Soc. Agr. Stat.

Cochran, W. G. (1963). Sampling Techniques, John Wiley & Sons, Inc., New York.

Durbin, J. (1959). A note on the application of Quenouille's Method of Bias Reduction in Estimation of Ratios. Biometrika 46, 477-80.

Hartley, H. O. and Ross, A. (1954). Unbiased Ratio Estimates, Nature, 174, 270-71.

Rao, J. N. K. and Webster, J. T. (1966). On two Methods of Bias Reduction in Estimation of Ratios. Biometrika, 53, 571-77

Srivastava, S. K. (1967). An Estimator Using Auxiliary Information in Sample Surveys. Calcutta Stat. Assoc. Bulletin, 16, 121-32.

Table 1: Efficiencies, E_1 and E_2, of t_1, (t_2 and t_3) over $\bar{y}$ and $\bar{y}_r$ for selected values of ρ and K and W = 1/4.

	K=0.5		K=1.0		K=1.5		K=2.0	
ρ	E_1	E_2	E_1	E_2	E_1	E_2	E_1	E_2
.1	101	116	99	178	94	277	89	400
.2	104	109	104	166	101	268	95	400
.3	106	101	110	153	109	257	105	400
.4	109	93	116	139	119	244	118	400
.5	112	84	123	123	131	229	133	400
.6	116	75	131	105	145	210	154	400
.7	119	65	140	84	162	187	182	400
.8	123	55	150	63	185	157	222	400
.9	126	44	163	33	215	118	285	400

Table 2: Efficiencies, E_1 and E_2, of t_1, (t_2 and t_3) over $\bar{y}$ and $\bar{y}_r$ for selected values of ρ and K and W = 1/2.

	K=0.5		K=1.0		K=1.5		K=2.0	
ρ	E_1	E_2	E_1	E_2	E_1	E_2	E_1	E_2
.1	99	114	87	157	71	209	56	256
.2	104	109	95	152	79	210	62	262
.3	110	104	105	147	90	211	71	271
.4	116	99	117	141	104	213	83	283
.5	123	92	133	133	123	215	100	300
.6	131	85	153	123	151	219	125	325
.7	140	77	182	109	195	224	167	367
.8	150	68	222	89	276	234	250	450
.9	163	57	286	57	471	259	500	700

Table 3: The exact efficiencies, E_r' and E_i', of $\bar{y}_r$ and t_i (i=1,2,&3) with W = 1/4, for selected values of m, K & ρ.

		ρ =.2				ρ =.3				ρ =.4			
m	K	E_r'	E_1'	E_2'	E_3'	E_r'	E_1'	E_2'	E_3'	E_r'	E_1'	E_2'	E_3'
8	.50	61	95	100	99	68	98	104	102	77	101	107	105
	1.00	37	93	96	98	43	99	102	104	51	106	109	111
	1.50	21	87	88	95	24	95	93	103	28	104	103	112
16	.50	77	100	103	101	86	103	105	104	96	106	108	107
	1.00	49	99	102	101	56	105	107	107	66	111	114	113
	1.50	29	94	96	98	32	102	104	106	37	112	114	116
20	.50	81	100	103	102	90	103	106	104	100	106	109	107
	1.00	51	100	102	102	59	106	108	107	69	112	114	114
	1.50	30	96	98	98	34	104	106	107	39	114	115	117
32	.50	86	102	103	102	95	104	106	105	107	107	109	108
	1.00	55	102	103	103	63	107	108	108	74	114	115	115
	1.50	39	98	99	99	37	106	107	108	43	116	117	117

Table 4: The exact efficiencies, E_r' and E_i', of $\bar{y}_r$ and t_i (i=1,2&3) with W = 1/2, for selected values of m, K & ρ.

		ρ =.5				ρ =.7				ρ =.9			
m	K	E_r'	E_1'	E_2'	E_3'	E_r'	E_1'	E_2'	E_3'	E_r'	E_1'	E_2'	E_3'
8	.25	79	94	99	99	86	99	103	105	91	103	105	113
	.50	87	104	109	109	117	120	126	126	168	140	146	149
	1.00	62	106	105	116	105	152	152	162	324	260	262	269
	1.50	33	87	78	102	50	139	122	163	103	340	264	415

Table 4: (continued)

m	K	ρ =.5				ρ =.7				ρ =.9			
		E'_r	E'_1	E'_2	E'_3	E'_r	E'_1	E'_2	E'_3	E'_r	E'_1	E'_2	E'_3
16	.25	100	103	108	106	111	109	114	112	123	115	120	120
	.50	109	114	119	116	147	130	136	113	222	152	157	156
	1.00	80	120	124	124	134	168	173	172	408	274	279	278
	1.50	44	105	107	112	67	167	168	179	139	409	391	444
20	.25	104	105	109	107	117	111	115	114	130	118	121	121
	.50	114	116	120	117	154	133	137	135	234	155	159	158
	1.00	83	123	127	126	140	171	175	174	425	277	280	279
	1.50	46	108	111	114	70	173	175	182	146	422	410	450
32	.25	111	108	110	109	125	114	117	116	142	121	124	123
	.50	121	119	121	120	164	136	138	138	252	158	161	160
	1.00	89	127	129	129	150	175	177	177	453	280	283	282
	1.50	50	114	116	118	76	181	183	187	159	441	435	458

Table 5: The absolute values of % Bias/$(MSE)^{1/2}$, B_r and B_i of $\bar{y}_r$ and t_i (i=1,2,&3) with W = 1/4, for selected values of m, K & ρ.

m	K	ρ =.2				ρ =.3				ρ =.4			
		B_r	B_1	B_2	B_3	B_r	B_1	B_2	B_3	B_r	B_1	B_2	B_3
8	.50	9.48	2.96	1.01	1.07	6.68	2.00	.69	.18	3.55	1.02	.35	.74
	1.00	19.74	7.80	2.64	3.92	18.57	7.05	2.38	3.12	17.29	6.24	2.11	2.26
	1.50	25.02	12.31	4.06	6.65	24.57	11.88	3.91	6.01	24.15	11.42	3.78	5.33
16	.50	7.03	2.00	.29	.74	4.94	1.35	.20	.10	2.62	.68	.10	.55
	1.00	14.90	5.31	.77	2.74	13.99	4.78	.69	2.16	12.99	4.22	.61	1.56
	1.50	18.56	8.42	1.22	4.67	18.22	8.10	1.17	4.21	17.91	7.77	1.12	3.72
20	.50	6.34	1.77	.20	.65	4.46	1.20	.13	.09	2.40	.61	.07	.50
	1.00	13.50	4.71	.53	2.45	12.66	4.42	.48	1.93	11.77	3.74	.42	1.39
	1.50	16.85	7.48	.84	4.17	16.54	7.12	.81	3.76	16.25	6.90	.77	3.32
32	.50	5.08	1.38	.09	.51	3.56	.93	.06	.06	1.88	.47	.03	.40
	1.00	10.86	3.68	.25	1.93	10.18	3.31	.22	1.52	9.43	2.92	.20	1.90
	1.50	13.62	5.86	.39	3.29	13.36	5.64	.38	2.96	13.12	5.40	.36	2.61

Table 6: The absolute values of % Bias/$(MSE)^{1/2}$, B_r and B_i of $\bar{y}_r$ and t_i (i=1,2,&3) with W = 1/2, for selected values of m, K and ρ.

m	K	ρ =.5				ρ =.7				ρ =.9			
		B_r	B_1	B_2	B_3	B_r	B_1	B_2	B_3	B_r	B_1	B_2	B_3
8	.25	8.99	4.89	1.67	5.71	16.86	9.04	3.09	9.71	25.11	13.36	4.49	13.98
	.50	0.00	0.00	0.00	2.29	8.75	4.42	1.51	6.62	20.97	9.57	3.25	11.73
	1.00	15.89	10.42	3.46	5.27	12.45	7.47	2.49	1.51	7.27	3.26	1.09	4.14
	1.50	23.15	18.80	5.95	12.10	22.88	19.04	5.94	10.58	24.55	22.36	6.56	9.29
16	.25	6.67	3.39	.50	4.09	12.65	6.27	.91	6.92	19.25	9.31	1.36	9.96
	.50	0.00	0.00	0.00	1.68	6.48	3.05	.44	4.75	15.88	6.58	.96	8.34
	1.00	11.89	7.31	1.06	3.67	9.26	5.18	.75	.96	5.38	2.21	.32	3.05
	1.50	17.64	13.65	1.96	8.62	17.42	13.79	1.98	7.46	18.79	16.17	2.26	6.35
20	.25	6.01	3.02	.34	3.67	11.44	5.59	.63	6.21	17.48	8.30	.94	8.92
	.50	0.00	0.00	0.00	1.51	5.85	2.71	.31	4.26	14.39	5.85	.66	7.47
	1.00	10.75	6.52	.74	3.27	8.37	4.61	.52	.83	4.85	1.96	.22	2.75
	1.50	16.00	12.25	1.38	7.71	15.80	12.38	1.38	6.67	17.06	14.50	1.59	5.64
32	.25	4.81	2.37	.16	2.91	9.19	4.39	.30	4.92	14.14	6.53	.44	7.06
	.50	0.00	0.00	0.00	1.21	4.68	2.12	.14	3.38	11.59	4.59	.31	5.92
	1.00	8.63	5.14	.35	2.57	6.70	3.62	.24	.64	3.88	1.53	.10	2.20
	1.50	12.92	9.74	.65	6.11	12.75	9.83	.66	5.26	13.79	11.49	.76	4.40

ACCESS TO HEALTH CARE: A PRELIMINARY MODEL

Martin K. Chen, Bureau of Health Services Research and Evaluation

Access to health care is an important concept both in theoretical discussions on health services and in empiric research on existing and new health care delivery systems. It is rare, however, for this author to find a precise definition of the concept in the literature. Usually administrators and researchers use the term to mean quite different things and the reader is left with the task of inferring its meaning in the particular context in which the term is used. For example, Fox (1), in a discussion on the Federal Government's role in increasing access to medical care for the poor, uses the term to mean ability to get into a health care system. Shannon et al (2) appear to use the term to mean "ready availability" of health services. Other investigators have used such diverse measures as patient travel time and utilization of services as indicators of access (3, 4).

The first attempt to define the concept known to this author is made by Given et al (5) in a paper that defines access as "the social, psychological, economic and organizational factors that influence individual participation in the health services system given the availability of services." Conceptually, this definition, like the World Health Organization definition of health (6), is comprehensive but vague. As such it practically defies operationalization. As a matter of fact, the authors themselves disregarded this definition later in the paper when they operationalized access as the ratio of the total number of doctor-patient contacts to the total number of disability days per 1000 population in the past two weeks.

A much more practical definition of access is the "use-need discrepancy ratio" used in the household survey conducted by the National Center for Health Services Research and Development (7). Symbolically, this ratio is defined as:

$$R = 100 \sum_{i=1}^{n} V_i \Big/ \sum_{i=1}^{n} R_i \text{ , where}$$

R = discrepancy ratio,

V_i = number of physician visits made by individual i for two-week period,

R_i = number of days of restricted activities, including bed days, within the two-week period for individual i, and

n = number of individuals included in the computation of R.

This definition is, in effect, another version of the operational definition by Given et al, and as such it shares the problems of the other definition. First of all, either of the two ratios is by itself uninterpretable without some kind of norm. For instance, if the discrepancy ratio for Population A is .33 (1/3) and that for Population B is .40 (2/5), this information in itself cannot help a health administrator decide whether he should try to increase access of care to either or both populations.

This is not to say, however, that the use-need discrepancy ratio is not useful. Given the resources and time, one could collect data from a large number of populations on disability days, physician contacts, and some indicator of health status. With such data, one could then statistically correlate the discrepancy ratio and health status. If the correlation is reasonably high, one could determine a point or range in the value of the discrepancy ratio that corresponds to the population with the highest health status. Without implying any causal relationship between the discrepancy ratio and health status, one could use the point or range as a norm in comparative studies.

Another problem has to do with sample size. Unless the sample size is substantial, the discrepancy ratio will have to be based on rather scanty data because of the low probability of people being bedridden or restricted during the past two weeks unless it is during the height of an epidemic season. This shortcoming is overcome to some degree when the time frame is a year rather than a two-week period, but the problem remains that a proportion of the sample surveyed do not contribute any information to the "discrepancy ratio."

In spite of their arbitrariness, the operational definition of Given et al and the "discrepancy ratio" make it possible to quantify a concept that has largely been left undefined. There is little doubt that a valid quantitative index of access, however imperfect, can be a very useful tool to both health program administrators and researchers interested in the evaluation of new or innovative health delivery systems. Toward this end, an attempt is made in this paper to operationally define and quantify access as a composite measure of several parameters useful in selected situations.

Before we define what access is, we need to differentiate what may be termed perceived access and access as objectively derived. These two concepts are different and may or may not be statistically correlated. It is useful to keep the two concepts apart because both are real in their effects on consumer behavior and a merger of the two may mask the dynamics of complex interaction patterns of intra-person and interpersonal factors vis-a-vis utilization of health services. In this paper we have chosen to focus on access as objectively derived, not because we believe perceived access unimportant, but because for quantification purposes perceived access requires a different type of data, such as obtainable by the questionnaire items on barriers to access quoted in Health Services Data System:

the Family Health Survey (8).

Attributes of a Useful Objective Index

For an objective index of access to be useful, it must meet the following conditions:

1. It must be based on data that are readily available or can easily be collected by a health delivery system;

2. It must possess invariance from region to region or situation to situation; in other words, the parameters of the index should be given identical definitions and the data collection procedures standardized across regions and/or situations;

3. It must be easily computable given the required data; and

4. It should not include subjective elements based on feelings or perceptions.

The desirability of Conditions 1, 2, and 3 is self-evident and need no amplification. The importance of Condition 4 may not be appreciated unless it is remembered that the proposed index is objective in nature, and as such it should not be contaminated with subjective data. Furthermore, when a subjective index of access is developed, it will then be possible to study the statistical relation of the two indices and to determine their separate and joint contributions, if any, to the variance of health service utilization of a given population or group.

Definition of Access

In this paper access is defined as the degree of difficulty of a potential user in getting into the health system, and once in the system, the degree of efficiency of patient handling, given that the potential user has the resources for service and that he or she appreciates the value of service. The two given conditions are intended to isolate the health delivery system, be it clinic or hospital, from two important personal factors that are external to the system and that are postulated to have an effect on utilization. By insulating the care-providing institution from these factors in our formulations we hope to allow administrators and researchers to focus attention on the institution itself vis-a-vis the problem of access.

As defined, access may be affected by either or both factors: (1) inadequacy of the system in terms of personnel, facilities and equipment, and services; and (2) lack of efficiency in the utilization of the available supply of personnel, facilities and equipment, and services. The proposed index incorporates information about both factors. If a health facility has an adequate supply of personnel, equipment and services but its access index is low, then it may be assumed that these resources are not efficiently utilized, and the administrator should examine the system to ascertain the causes of the lack of efficiency.

Accordingly, the proposed index consists of two components, one component pertaining to a measure of adequacy of the physical facilities and personnel of the institution and the other component to efficiency in patient admission and handling. Let C_1 represent the first component. Then:

$$C_1 = \sum_{j=1}^{3} v_j \sum_{i=1}^{n_j} w_i \left| (I_i - R_i) \right| , \qquad (1),$$

where: I_i = ideal number of types of service, personnel or equipment; R_i = actual number of types of service, personnel or equipment; w_i = weight for the absolute difference between the ideal number and the actual number; v_j = weight for a given category (i.e., services, personnel or equipment); $i = 1,2, \ldots . n_j$; and $j=1, 2, 3$.

It is seen from Equation (1) that C_1 is a simple linear function of the absolute differences between the ideal number and the actual number for each type of service, personnel and equipment weighted in some manner. Although the absolute differences are used, positive and negative differences should be given different weights. This is so because the consequences of a positive difference (i.e., the ideal number exceeds the actual number) cannot be the same as those of a negative difference (i.e., the actual number exceeds the ideal number). For example, if the ideal number of ambulances for a service area is four, but the actual number is only two, many lives in the area may be threatened. This is not true if the numbers are reversed, although the situation is economically undesirable. The values of v_j reflect the relative importance of deficiencies in services, personnel and equipment.

If the economic factor of over-supply in equipment, services and personnel is ignored, Equation (1) can be reduced to:

$$C_1 = \sum_{j=1}^{3} v_j \sum_{i=1}^{n_j} w_i \, (I_i - R_i) \qquad (2)$$

Equation (2) is identical with Equation (1) except for the absolute sign. With Equation (2), $(I_i - R_i)$ is set to zero whenever its value is negative. Negative values are not allowed in the equation because they may neutralize the positive values, thus obscuring the different areas of deficiency in the system.

As formulated, C_1 is inversely related to adequacy; that is, the higher the value of C_1, the less adequate the supply of personnel, services and equipment of the institution. A computational example based on Equation (1) and the data from Tables 1 and 2 is given below:

For Facility 1,

$$\begin{aligned} C_1 = {} & (.4)[5(0)+3(1)+1(2)+1(1)] \\ & + (.2)[2(4)+2(3)+4(1)] \\ & + (.4)[3(1)+3(2)+3(2)] \\ = {} & (.4)(6)+(.2)(18)+(.4)(15) = 12.0 \end{aligned}$$

For Facility 2,

$$C_1 = (.4)[2(1)+5(0)+2(1)] + (.2)[3(6)+3(2)+1(3)] + (.4)[5(4)+4(0)+2(2)+1(1)] = (.4)(4)+(.2)(27)+(.4)(25) = 17.0$$

It is noted that the lowest possible value of C_1 is zero, indicating exact correspondence between the ideal setting and the actual setting in terms of health services, personnel, and equipment. As the value of C_1 goes up, the deviation of the actual setting from the ideal becomes greater. Since the value of C_1 cannot be negative, its value cannot reflect the direction of the deviation; for that information one is referred to Table 2, where positive and negative differences are given.

Efficiency of Patient Handling

Now let C_2 be the second component of the index representing the degree of efficiency in patient admission and handling. Then

$$C_2 = f(A, T, W, P) \qquad (3),$$

where A is appointment waiting-time in days, T is patient traveling time to the care delivery institution in minutes, W is waiting-room time in minutes, and P is throughput time from first contact with physician or other health professional to completion of visit, also in minutes. In words, this component of the index is a function of four parameters, all of which have to do with the duration of the patient visit. The smaller the value of any of the four parameters the shorter the duration of the visit and the more efficient the patient handling.

Intuitively, C_2 should be some kind of average of the four parameters. Since T, W, and P are in units of minutes and A is in units of days, we make the units commensurate by transforming an eight-hour day into minutes by simple multiplication, or 60 x 8 = 480. Then we derive C_2 as:

$$C_2 = \left[\left(\frac{a_1 \sum_i^n T_i/n + a_2 \sum_i^n W_i/n + a_3 \sum_i^n P_i/n}{a_1 + a_2 + a_3} \right) \cdot 480 \sum_i^n A_i/n \right]^{1/2}$$

$$= \left[\left(a_1 \sum_i^n T_i/n + a_2 \sum_i^n W_i/n + a_3 \sum_i^n P_i/n \right) \cdot 480 \sum_i^n A_i/n \right]^{1/2}$$

$$= \left(\bar{X}_1 \cdot 480\, \bar{X}_2 \right)^{1/2} \qquad (4)$$

where a_1, a_2 and a_3 are weights; $a_1 + a_2 + a_3 = 1$; and n is the number of patients sampled from a facility. We use the subscript i to represent sample patients in any facility and the n's need not be equal across facilities. It must be remembered, however, that the means of small n's are not reliable. Further, the weights for the different parameters must be the same across facilities to ensure invariance.

C_2 is actually two types of averages. The first term,

$$(1/n)\left(a_1 \sum_i^n T_i + a_2 \sum_i^n W_i + a_3 \sum_i^n P_i\right)$$

is a weighted arithmetic mean of the values of T, W, and P for all sampled patients and the second term,

$$\left(480 \sum_i^n A_i/n \right)$$

is the arithmetic mean of the values of A for all patients weighted by a constant. C_2 is simply the geometric mean of the two terms. The reason the geometric mean of the two terms is used rather than the arithmetic mean is that the two quantities are expected to be quite disparate, and when this is the case, the arithmetic mean tends to be automatically weighted toward the larger quantity and give negligible weight to the smaller quantity. For instance, the arithmetic mean of 1 and 100 is 50.5, but the geometric mean is 10. The two means will approximate each other as the two numbers approximate each other in magnitude, and will be identical if the two numbers are equal.

Mathematically, this phenomenon is explained by the fact that in using the geometric mean we have in effect done a logarithmic transformation of the original quantities to reduce their distance. In terms of logarithms, Equation (4) is written as:

$$\log C_2 = \log \left(\bar{X}_1 \cdot 480 \bar{X}_2\right)^{1/2} = (1/2)(\log \bar{X}_1 + \log 480 + \log \bar{X}_2) = (1/2)(\log \bar{X}_1 + 2.68 + \log \bar{X}_2) \qquad (5)$$

It is seen from Equation (5) that the second term, $480X_2$, becomes two additive quantities, the constant 480 now being 2.68. This property of logarithmic transformations makes them a valuable tool in dealing with averages of numbers that are widely disparate in magnitude.

Examination of Table 3 reveals several things worthy of note. First, regardless of which of two sets of weights is used, C_2 is larger for Facility 2 than it is for Facility 1, indicating that Facility 1 is more efficient than Facility 2. Second, it is seen that when a new set of weights is used, the value of C_2 tends to increase in Facility 2, but decrease in Facility 1, although there is no change in the rank order of the two facilities.

Index of Access

With the values of C_1 and C_2 known or computable, we now propose an index of access:

$$I_x = \arcsin(10^k C_1/N)^{1/2} + \arcsin(10^r C_2/N)^{1/2} \qquad (6),$$

where I_x is the index of access, k and r are single-digit integers, and N is the population of users of the facility in a catchment area.

Equation (6) is actually an angular transformation of two ratios, the first being the ratio of the number of deficiencies to the size of the population and the second the ratio of the number of wasted minutes to the same population. This

transformation serves two functions; namely, to make the two components of the index additive and to stabilize the variances of the two ratios because of the known statistical relation between the mean and the variance of ratios or proportions. The constants k and r are intended to adjust the values of the ratios such that the ratios are not too small or too large. Their values range between -9 and +9. In any comparative study, the constants must have identical values across different facilities.

Computationally, Equation (6) may look complex, but in reality it is simple. A table such as Table 4 with given values of C_1 and C_2 as well as N would help. The values of k and r can be easily assigned after looking at the ratios of C_1 or C_2 to N. In this case it is seen that if we add two zeroes to C_1 and one zero to C_2, the ratios should be just right. Accordingly, we give the value of 2 to k and the value of 1 to r. Once the ratios are computed, we take their square roots, which can be done on a desk calculator or by referring to a table of square roots. Then the square roots are converted to degrees by using a standard table available in most statistical textbooks.

I_x has an inverse relationship with access; that is to say, the higher the value of I, the less accessible the facility is. Table 4 shows that the I_x value for Facility 1 is 97.98 and that for Facility 2 is 115.70, indicating that Facility 1 is more accessible than Facility 2. Note that I_x is a function of three parameters, C_1, C_2 and N. By holding two of the parameters constant, it is seen that the value of I_x will be high if either C_1 or C_2 is high. On the other hand, the value of I_x will be low if N is high. This phenomenon makes sense in that if a facility can keep the values of C_1 and C_2 comparatively small while it has a larger population to serve, it is bound to be more accessible than another facility with the same values of C_1 and C_2, but with a smaller population to serve.

Weighting Problems

For computing both C_1 and C_2, basic components of the index, weights are assigned to the various parameters. What should these weights be? In the case of C_1, two different weighting scales are used, the v_j's and w_i's. The weights are entirely arbitrary for demonstration purposes. Where external criterion or criteria are available, it is theoretically possible to collect enough data to statistically determine the optimal weights for the different categories of personnel, equipment and services to minimize the error of predicting the criteria by a linear combination of these categories. For example, the degree of adequacy of personnel, equipment and services may be statistically related to patient satisfaction as an external criterion. If so, different least squares techniques can be used to determine the optimal weights for personnel, equipment and services such that the error of predicting patient satisfaction by a linear combination of the three parameters is minimum.

In the case of C_2, the relative importance of T or travelling time, W or waiting-room time, and P or patient processing time, must be somehow determined. Here again, patient satisfaction may be used as an external criterion to be predicted by a linear combination of the three parameters such that the error of prediction is minimum. The weights so determined will be objective based on available data.

There are two basic problems with this approach. One is that cumulation of the data necessary for least squares solutions, if at all feasible, will be time-consuming because of the lack of valid instruments for the measurement of external criterion, such as patient satisfaction. The developmental work alone may take six months or longer. Validating the instruments may require another six months.

Another problem is that, even if a valid external criterion were readily available, the weights for the different parameters determined on the basis of data from one sample of patients would have to be cross validated with data from other samples. The patient population of a health care institution may change character over time, thus aggravating the already serious problem of sampling variations of the least squares weights. Furthermore, once a stable set of weights has been determined, periodic updating and perhaps revision with new data must be undertaken to ensure the continued validity of the weights.

An alternative to the statistical technique of minimizing the error of prediction is the psychometric technique of scaling. The values of the weights for the different parameters, say T, W, and P for C_2, can be scaled by a variety of techniques, such as pair comparison and successive categories discussed in Guilford's work (9). With proper sampling procedure, the scale values so derived represent the values to which a given population subscribes. Although still subjective in nature, the weights or values so determined may in fact have greater intrinsic validity than objective weights because it is well known that people's behavior is governed by their perceptions of reality than by reality per se.

The above discussions may give the impression that the utility of the model depends on the proper determination of the weights in C_1 and C_2. That is not true. Initially, one could give the parameters equal weights by assigning the value of one to them. Or, one could use a panel of judges to determine the rank order of the parameters and use the ranks as weights. The comparisons of the health care institutions in terms of access will be valid if the weights are accepted by these institutions.

Some Cautionary Remarks

It is stated earlier in this paper that the index should be useful in selected situations, implying that certain requirements must be met for the valid use of the index. What are these requirements? One is that the care providing institutions have available data or are willing to

collect such data on new patients or old patients seeking appointments of their own accord. This requirement is intended to rule out physician-ordered or recall visits that are prescheduled and that would make appointment waiting time meaningless. If the physician orders a patient to return for a checkup in three months, it is not fair to the physician or clinic to use three months as A or appointment waiting time.

Another requirement is that the data are limited to non-emergency cases. In true emergency cases there usually is no appointment necessary and appointment waiting time is the time interval between the call to the ambulance service and the arrival of the ambulance, probably a matter of minutes. Since true emergency cases constitute only a small fraction of the total caseload of an institution, it is better not to apply the index to these cases.

The third requirement is dictated as much by validity as by common sense. That is, for comparative purposes the institutions should be similar in nature and serving similar types of populations. One would not compare a fee-for-service multi-specialty group with a public-supported hospital, such as a U.S. Public Health Service hospital. They are different in nature and they serve different types of populations. While the question, "How similar is similar?" may be legitimately raised, one need not be slavish in matching the institutions to be compared or the populations to be served. Nonetheless, the closer the match of the institutions compared, the easier it is to pinpoint the causes of their relative efficiency as measured by I_x.

The fourth and last requirement is that for each institution there is a well-defined catchment area. Without a well-defined area, it would be futile to talk about "ideal numbers" of personnel, equipment and services. The numbers are "ideal" in relation to the population served; if this population were unknown or only vaguely known, then it would be impossible to come up with an ideal that has any meaning. This requirement may be difficult to meet because of the plurality of service institutions in a community, except for prepaid group practices.

The first component of the index, C_1, deals with quantity only. An additional weight for quality could easily be incorporated into the formula, but this would make the index computationally complex. Besides, quality, particularly of professional people, is difficult to measure. For this reason the parameter of quality is ignored in the present model.

Finally, it is to be remembered that this index is not a precise indicator of access. It is some number that has no meaning in itself, but that it assumes meaning only when the same set of operations are applied to selected institutions that meet certain requirements.

REFERENCES

1. Fox, Peter D., "Access to Medical Care for the Poor: The Federal Perspective," Medicare Care 10:272-277, May-June 1972

2. Shannon, G. W., L. Bachshur, and C. A. Metzner, "The Concept of Distance as a Factor in Accessibility and Utilization of Health Care," Medical Care Review 26:143-161, 1969

3. Drosness, D. L. and J. W. Lubin, "Planning Can Be Based on Patient Travel," The Modern Hospital 106:92-94, April 1966

4. Willie, C. V., "Health Care Needs of the Disadvantaged in Rural-Urban Area," HSMHA Health Reports 87:81-86, January 1972

5. Given, C., C. Akpom, and S. Katz, "Management Information for Improvement of Access to Ambulatory Care Facilities," unpublished report

6. World Health Organization, Measurement of Levels of Health, Technical Report Series No. 130, Geneva, World Health Organization, 1957

7. National Center for Health Services Research and Development, Health Services Data System: The Household Survey, unpublished report, August 1972

8. Eichhorn, R. L., Health Services Data System: The Family Health Survey, unpublished report

9. Guilford, J. P., Psychometric Methods, New York:McGraw-Hill Book Co., 1954

Table 1

Values of w_j and v_j for the Computation of C_1 for Two Health Facilities*

Facility 1					Facility 2				
I. Types of Service (v_1 = .4)	1	2	3	4	I. Types of Service (v_1 = .4)	1	2	3	
$w_i(+)$	5	3	3	2	$w_i(+)$	2	5	3	
$w_i(-)$	2	2	1	1	$w_i(-)$	1	3	2	
II. Types of Personnel (v_2 = .2)	1	2	3		II. Types of Personnel (v_2 = .2)	1	2	3	
$w_i(+)$	2	4	4		$w_i(+)$	3	3	2	
$w_i(-)$	1	2	3		$w_i(-)$	2	2	1	
III. Types of Equipment (v_3 = .4)	1	2	3		III. Types of Equipment (v_3 = .4)	1	2	3	4
$w_i(+)$	3	3	3		$w_i(+)$	5	4	2	2
$w_i(-)$	2	2	1		$w_i(-)$	3	2	1	1

* For v_j a scale from 0 to 1 is used, whereas for w_j a scale from 1 to 5 is used. The scales are entirely arbitrary. $w_i(+)$ refer to weights to be used if (I_i-R_i) is positive, and $w_i(-)$ to weights to be used if (I_i-R_i) is negative.

Table 2

Comparison of Two Health Facilities in Access With Fictitious Data

Facility 1					Facility 2				
I. Types of Service	1	2	3	4	I. Types of Service	1	2	3	
Ideal (I_i)	1	3	2	0	(I_i)	4	3	7	
Actual (R_i)	1	2	4	1	(R_i)	3	3	8	
Diff.	0	+1	-2	-1		+1	0	-1	
II. Types of Personnel	1	2	3		II. Types of Personnel	1	2	3	
Ideal (I_i)	20	12	8		(I_i)	24	15	5	
Actual (R_i)	16	15	7		(R_i)	18	13	8	
Diff.	+4	-3	+1			+6	+2	-3	
III. Types of Equipment	1	2	3		III. Types of Equipment	1	2	3	4
Ideal (I_i)	3	2	5		(I_i)	7	2	5	0
Actual (R_i)	2	0	3		(R_i)	3	2	3	1
Diff.	+1	+2	+2			+4	0	+2	-1

Table 3 [1/]

Hypothetical Data for Computing C_2

	Facility 1				Facility 2			
User	T	W	P	A	T	W	P	A
1	5	12	30	4.5	10	21	8	7
2	13	5	26	6	7	15	10	3.5
3	17	20	15	3	20	21	30	6
4	32	4	7	8	23	20	28	7.5
5	12	15	25	4	18	19	15	9.5
6	10	8	11	10				
Total	89	64	114	35.5	78	96	91	33.5
Mean	14.8	10.6	19	5.9	15.6	19.2	18.2	6.7
C_2* = 207.6			C_2** = 198.8		C_2* = 238.7		C_2** = 241.8	

* Computations based on following weights: $a_1 = .3$, $a_2 = .3$, $a_3 = .4$

** Computations based on following weights: $a_1 = .2$, $a_2 = .5$, $a_3 = .3$

1/ To facilitate computation, a table such as Table 3 may be used.

Table 4

Computational Table for Comparing Two Facilities in Terms of Access

	Facility 1	Facility 2
C_1	12.0	17.0
C_2	207.6	238.7
N	5000.0	4000.0
k	2	2
r	1	1
I_x	97.98	115.70

EQUITY AND EFFICIENCY IN STATE AID TO PUBLIC SCHOOLS*

Elchanan Cohn, The Pennsylvania State University

I. INTRODUCTION

Legal and legislative battles concerning state aid to education have placed the issue of state aid at or near the top of the list of priorities regarding public education. Already several states have adopted new state aid schemes, and a move is afoot to reform other existing state aid formulas.

Until very recently, the most common method of state financing of public schools has been based on the so-called foundation program. Several states have recently adopted a variant of the foundation program which is known as the percentage equalizing plan. Other states have used another variant of the foundation program which is known as the guaranteed valuation or resource equalizer plan. The newest breed of state aid schemes is the district power equalization plan, which is supposed to insure that educational funds raised by any district will be entirely unrelated to community wealth.[1]

There is a degree of equalization in all of the aid formulas. However, sufficient evidence has been presented to indicate that the foundation or percentage equalizing approaches do not eliminate considerable variation in educational expenditures by school districts. Community wealth remains an important determinant of a district's ability to raise educational funds. Whether such a situation is unconstitutional or simply undesirable is a matter that should be left to the courts or the political decision-making process, respectively. What is of concern to economists is the degree to which a given state aid scheme is shown to result in greater equalization or other outcomes (such as a reduction in the local property tax burden).

To gain a measure of understanding of the effect of state aid on local revenues and expenditures for education, a number of scholars have studied the determinants of educational expenditures. Some studies employed single-equation models, combining supply and demand variables in a single equation (examples are Miner [1963], Brazer [1959], Renshaw [1960], Bishop [1964], and Sacks [1972]). Other studies have attempted to describe supply and demand structures for educational expenditures (examples are McMahon [1970] and Booms and Hu [1971]). Although the conclusions differ from study to study, the majority of studies indicate a regression coefficient for state aid between 0 and 1, suggesting that state aid is both substitutive (some of the state aid money is used for other public goods or for a reduction in the tax burden) and stimulative (*total* educational expenditures will increase due to state aid).

In Section II, the effect of state aid on educational expenditures and revenue will be studied along with its effect on three other variables: nonpublic enrollment rates, average school size, and bond sales.

II. A CROSS-SECTIONAL, INTERSTATE MODEL: 1967-68

The empirical model presented here provides additional insights regarding the effect of state aid on educational expenditures, employing both a new structure and more recent data. The model also provides a first attempt to study the effect of state aid on three variables: average school size, nonpublic enrollment rates, and bond sales. A fifth variable to be studied is local revenues.

In addition to the state aid variables, each of the variables to be investigated here is also a function of other factors. First, some of the (endogenous) variables mentioned above might influence one another. For instance, per pupil expenditures in a given state are likely to be a function of school size, as several studies (to be discussed below) have indicated. Or, local revenues may be a function of the percent of enrollment in nonpublic schools. Furthermore, other (exogenous) factors may influence the variables under investigation. For example, the degree of urbanization in the state is likely to affect average school size, local revenues, and per pupil expenditures. Local revenues and expenditures may also be affected by the perceived "quality" of the public schools. Two measures of "quality" are average teachers' salaries and the student/teacher ratio.

The Model

Let $Y_1, Y_2, \ldots, Y_5$ denote the endogenous variables, and $X_1, X_2, \ldots, X_{10}$ the exogenous variables. Both variable sets are defined in Table 1. The empirical model is given in Equations (1) through (5).

$$(1)\quad Y_1 = f_1(Y_3, Y_4, Y_5;\ X_1, X_2, X_6, X_7 X_8)$$

$$(2)\quad Y_2 = f_2(Y_1, Y_3, Y_4;\ X_1, X_4, X_5, X_6, X_7, X_{10})$$

$$(3)\quad Y_3 = f_3(Y_1, Y_2, Y_5;\ X_1, X_3, X_5, X_6, X_7, X_9, X_{10})$$

$$(4)\quad Y_4 = f_4(Y_2, Y_3, Y_5;\ X_1, X_3, X_5, X_6, X_7, X_{10})$$

$$(5)\quad Y_5 = f_5(Y_1, Y_3, Y_4;\ X_1, X_4, X_5, X_6, X_7, X_{10})$$

A linear form is assumed for each equation.

It is hypothesized in Equation (1) that the larger the percentage of pupils enrolled in nonpublic schools, the smaller would the average school size be, other things equal. It also appears plausible that the variable Y_4 should be related to school size, but there are two conflicting forces; on the one hand, if proceeds from bond elections are used to build larger schools, the effect on relative size would be positive; on the other hand, if such proceeds are used to reduce crowding by building additional

---

*This paper is based on the author's *ECONOMICS OF STATE AID TO EDUCATION* (Lexington, Mass.: Lexington Books, D.C. Heath & Company, in press). The study was supported by a grant from the National Institute of Education. Points of view stated herein do not necessarily reflect official position or policy of the National Institute of Education.

[1] For a discussion of these and other plans, consult Cohn (1974).

TABLE 1
MEANS, STANDARD DEVIATIONS, DEFINITIONS, AND SOURCES OF VARIABLES

Variable	Mean	Standard Deviation	Definitions of Variables
Endogenous			
Y_1	392.59	144.18	Relative size of schools (pupils in ADA per school) 1967-1968.
Y_2	$625.48	125.83	Current expenditures per pupil in ADA (Average Daily Attendance), 1967-1968.
Y_3	0.10	0.061	Percent of pupils enrolled in nonpublic schools, 1967-1968.
Y_4	$465.99	364.64	Total approved par value of bond issues, 1962-1971, per pupil enrolled in public elementary and secondary schools.
Y_5	$379.60	152.26	Local revenue per pupil, 1967-1968.
Exogenous			
X_1	$275.41	111.42	State aid per pupil in ADA, 1967-1968.
X_2	23.09	2.12	Percent of total population enrolled in public schools, 1967-1968.
X_3	$2,955.10	506.12	Personal income per capita, 1967.
X_4	$13,999.59	3,348.94	Personal income per pupil in ADA, 1967.
X_5	5.07	1.12	Equalization score of state, 1968-1969.
X_6	11.74	12.21	Negro enrollment in public schools as a percent of total enrollment, 1968.
X_7	65.42	14.44	Urban population as a percent of total population, 1970.
X_8	13.36	5.57	Incidence of poverty, 1969 (percentage points).
X_9	$7,161.59	1,025.38	Average teachers' salary, 1967-1968.
X_{10}	0.023	0.0019	Number of students per 1,000 teachers, 1967-1968.

SOURCES:

1. Richard H. Barr and Geraldine J. Scott, *STATISTICS OF STATE SCHOOL SYSTEMS, 1967-1968* (Washington, D.C.: U.S. Office of Education, 1970) -- for the following variables: Y_1, Y_2, Y_5, $X_1 - X_4$, X_9, X_{10}.

2. Roe L. Johns and Richard G. Salmon, "The Financial Equalization of Public Support Programs in the United States for the Year 1968-1969," in *STATUS AND IMPACT OF EDUCATIONAL FINANCE PROGRAMS*, Vol. 4, ed. by Roe L. Johns, et al. (Gainesville, Florida: National Educational Finance Project, 1971), p. 137-- for X_5.

3. U.S. Bureau of the Census, *STATISTICAL ABSTRACT OF THE UNITED STATES: 1969, 1970, and 1971 EDITIONS* (Washington, D.C.: Government Printing Office, 1969, 1970 and 1971) -- for Y_3, X_6-X_8.

4. Irene A. King, *BOND SALES FOR PUBLIC SCHOOL PURPOSES* (Washington, D.C.: U. S. Office of Education, 1972) -- for Y_4.

schools (not necessarily of larger average size), then the effect on average school size might be negative. For the same reason, it is not clear *a priori* how Y_5 and Y_1 are related.

Among the exogenous variables in the set, five were included in the equation. For state aid, a negative coefficient is expected, as additional state aid might reduce incentives for school reorganization. The variable X_2 (percent of population enrolled in public schools) indicates the relative demand for public educational facilities in the state. The greater the demand, the greater the average school size is expected to be, other things equal. It is further expected that school size will be directly related to the percentage of Negro enrollment because of the observed overcrowding in areas where large concentrations of Negroes exist. Also, because urban areas are likely to have far greater population densities, greater urbanization should be positively related to school size, other factors remaining the same. Finally, the variable X_8 has been added to the equation to account for the expected negative relationship between school size and poverty in states where considerable rural poverty exists.

Concerning Equation (2), the determinants of expenditures include three endogenous and six exogenous variables. Because scale economies are expected to occur in public school operations, the hypothesized relationship between Y_1 and Y_2 is negative.[2] (A parabolic relationship, indicating a U-shaped relation between the two variables, was found to be nonsignificant; hence, only the linear term has been left in the equation.) It is also hypothesized that the greater the percentage of pupils enrolled in nonpublic schools, the higher would Y_2 be because local educational revenues collected from all citizens without regard to school enrollment would be distributed over a relatively smaller student population. Furthermore, it is expected that higher values of Y_4 would be directly correlated with Y_2 because the variable Y_4 is indicative of the citizens' attitude toward education. If they are willing to approve bond issues, they would probably also desire higher per pupil expenditures.

The variable X_4 is included in the equation to account for differences in wealth per pupil among states. It is hypothesized that a higher equalization score would be commensurate with higher per pupil expenditures, that expenditures are lower in states with large Negro enrollments but higher in urban areas, and that greater school quality requires more expenditures, so that X_{10} and Y_2 should be negatively correlated. A positive coefficient for X_1 is expected.

Three endogenous and seven exogenous variables are included in Equation (3). It is hypothesized that as school size increases, especially because of overcrowding, more parents will send their children to private schools. But if per pupil expenditures are greater, fewer parents will seek private education for their children. The effect of Y_5 on Y_3 is not unambiguously clear. On the one hand, more local revenues imply more local expenditures, with the likelihood that greater quality in public schools would encourage parents to send their children to public schools. However, if Y_5 is directly related to community wealth, the relationship between Y_5 and Y_3 might be positive. It is possible, of course, that Y_5 might be greater not because of greater wealth but because of greater tax effort, implying a more favorable attitude toward--and therefore greater rates of attendance in--public education.

Since X_3 provides a measure of average wealth it is expected to be directly related to nonpublic enrollment rates. It is also hypothesized that greater equalization would lead to greater nonpublic enrollments, as would be the case for greater levels of the variables X_6 and X_7. On the other hand, greater school "quality" in the form of higher salaries or *lower* student/teacher ratios should be negatively related to private enrollment rates. The *a priori* effect of state aid is not clear: on the one hand, if more state aid is synonymous with greater *equalization* the effect on Y_3 might be positive. On the other hand, if more state aid is synonymous with greater educational *quality*, the coefficient might be negative. Hence, no *a priori* expectations are stated in this case.

Three endogenous and six exogenous variables form the specification of Equation (4). It is hypothesized that Y_2 is indicative of a community's attitude toward support of public education; hence, a direct relationship between Y_2 and Y_4 is anticipated. Conversely, if a greater proportion of pupils attend nonpublic schools, parents would be more reluctant to support the public schools. It also appears that greater local revenues imply less need for bond financing. However, since Y_5 could also be a proxy for local capacity to absorb the financing of the bond as well as community's attitude, it is not clear what sort of relationship one should expect between Y_5 and Y_4.

If per capita income (X_3) is indicative of a community's attitudes, a positive correlation between X_3 and Y_4 would be expected. Such a relationship would be strengthened when it is recognized that wealthier communities are likely to be able to absorb the cost of bond financing with relatively greater ease than is the case in poorer districts. On the other hand, it is expected that a higher value of X_5 would result in lower bond sales since incentives for long-term indebtedness by local governments are reduced. Moreover, because of the general deterioration of the urban areas in the United States, especially in cities where the percentage of nonwhite population is relatively large, it is expected that a negative correlation between X_6 and Y_4, as well as between X_7 and Y_4, will be found. Since a smaller student/teacher ratio requires more facilities, a negative relationship between X_{10} and Y_4 is expected. Finally, since state aid could be substituted for local financing, a negative coefficient for X_1 is hypothesized.

Three endogenous and six exogenous variables have been included in Equation (5). The first hypothesis is that because of anticipated scale economies, greater school size would be negatively related to local revenue requirements,

[2] For studies on scale economies, consult, for example, Cohn (1968), Cohn and Hu (1973), Riew (1966), and Sabulao and Hickrod (1971).

other things equal. The effect of Y_3 on Y_5 is not unambiguously clear. On the one hand, higher private enrollment rates indicate unfavorable attitudes toward the public schools, pointing to a smaller level of Y_5. On the other hand, states with higher private enrollment rates may also be associated with relatively wealthier districts, in which case revenues for an equal tax effort should be greater. A positive sign is expected for Y_4 for two reasons. First, the variable is indicative of community attitudes. Second, a greater value for Y_4 is also indicative of greater debt service requirement, which should increase the demand for local revenues.

Per pupil income, as a measure of wealth, should be positively correlated with Y_5. But X_5 is hypothesized to be negatively correlated with Y_5 because greater equalization is expected to reduce the incentives of many school districts to raise revenues from local sources. It is hypothesized that local revenues in areas with higher levels of the variables X_6 and X_7 would be smaller and that greater school "quality," measured by X_{10}, would require greater local revenues; hence, X_{10} and Y_5 should be negatively correlated. Finally, since the literature review produced both positive and negative coefficients for the effect of state aid on local expenditures, no *a priori* hypothesis is advanced in this case.

Data

To implement the model, data have been assembled from various sources, principally publications of the United States Office of Education. The unit of observation is the state, and data are available for forty-nine states. (Hawaii has been excluded because it is essentially one large school district and therefore is not suitable for the present analysis.) The definitions of the variables used in this study--along with some descriptive statistics--are provided in Table 1.

Although the data are (with exceptions) for the year 1967-68 and hence do not portray the *current* state of affairs in public education, the relationships which we seek to derive are probably as relevant today as they were during the 1967-68 period--and this despite the changes that have occurred since that period in educational finance and administration.

Regression Results

The regression results are reported in Table 2. For each of the Equations (1) through (5), the table reports the coefficients derived on the basis of the Two State Least Squares (TSLS) estimation procedure (that is, when Equations (1) through (5) are considered as a system of equations, and the coefficients in Equation (1') through (5') account for the interdependence among the equations).

Average School Size: The Interstate data explain almost 80 percent of the variations in average school size. Contrary to hypothesis, state aid appears to contribute positively to that variable. Since our study of the state aid formulas showed little, if any, incentives for attaining optimal school size, it is difficult to conclude that more state aid is the *cause* of larger school size. A possible explanation of the positive correlation is that states that happen to have larger schools are the ones that also happen to give more aid to local districts. Nevertheless, the negative correlation that we expected was definitely refuted by the data.

Concerning the other explanatory variables, five variables are statistically significant. As hypothesized, the sign of the coefficients of both X_6 and X_7 is positive, and the sign of X_8 is negative. Also, the results suggest that, as expected, when enrollments in nonpublic schools are greater, average school size is likely to be smaller. On the other hand, contrary to expectations, the data indicate that a greater relative demand for education, measured by the percentage of total population enrolled in public schools, is associated with smaller school size.

Expenditures Per Pupil: The data confirm the expected relationship between state aid and expenditures. For each $1.00 of state aid, expenditures per pupil are likely to increase by $0.36. The coefficient is statistically significant at 0.10 level. These results suggest that state aid is likely to be both stimulative and substitutive: on the one hand, more state aid implies higher expenditures (stimulative); on the other hand, the results suggest that *local* expenditures are reduced by $0.64 for each $1.00 of state aid.

Among the remaining explanatory variables in Equation (2'), the only variable that has a relatively large t-ratio (significant at the 0.10 level) is Y_4, suggesting a positive net correlation between expenditures and bond sales.

Nonpublic Enrollment Rates: Three variables are significant at the 0.05 level: Y_1, X_1, and X_9. The coefficient of Y_2 is significant at the 0.10 level. The coefficient of X_1 is *negative*, and the signs of the coefficients of Y_1 and X_9 are consistent with *a priori* expectations. The negative sign of X_1 provides a measure of credence to the hypothesis that state aid has a lesser impact on equalization than on overall improvement in the quality of education.

Approved Value of Bond Issues: The coefficient of X_1 is significantly negative at the 0.10 level, indicating lower bond sales in states where higher state aid is given. This is consistent with our *a priori* expectations. The only other significant variable is Y_3, which has a negative coefficient. This is consistent with recent reports of school bond election results in Detroit and other areas with large nonpublic enrollments.

Local Revenue: State aid (X_1) is the only variable with a statistically significant coefficient. The negative sign of the coefficient indicates that, on the average, some substitution of state for local funds takes place.

III. CONCLUSIONS

The model provides several insights into the economic effects of state aid. With the exception of average school size, our *a priori* expectations of such effects were confirmed by the analysis. The results indicate that a greater level of state aid is associated with greater per

TABLE 2
REGRESSION RESULTS
(TWO-STAGE LEAST SQUARES)

(1') $Y_1 = 1097.66 - \underset{(2.23)}{2{,}018.28}Y_3 - \underset{(0.34)}{0.018}Y_4 + \underset{(0.56)}{0.22}Y_5 + \underset{(3.33)}{0.40}X_1 - \underset{(19.47)}{41.91}X_2 + \underset{(1.87)}{4.20}X_6 + \underset{(5.97)}{6.40}X_7 - \underset{(2.02)}{7.60}X_8$

$\bar{R}^2 = 0.79$, SEE = 65.86, F = 27.00

(2') $Y_2 = 761.15 + \underset{(0.70)}{0.28}Y_1 + \underset{(0.90)}{1{,}347.14}Y_3 + \underset{(1.87)}{0.23}Y_4 + \underset{(1.70)}{0.36}X_1 - \underset{(0.24)}{0.006}X_4 - \underset{(0.97)}{13.38}X_5 - \underset{(0.40)}{1.03}X_6 - \underset{(0.23)}{0.56}X_7 - \underset{(1.26)}{16{,}420.16}X_{10}$

$\bar{R}^2 = 0.88$, SEE = 42.69, F = 41.99

(3') $Y_3 = 0.22 + \underset{(2.18)}{0.00089}Y_1 + \underset{(1.95)}{0.00077}Y_2 + \underset{(0.01)}{17.86}Y_5 - \underset{(2.57)}{0.00048}X_1 + \underset{(0.55)}{0.000035}X_3 + \underset{(0.85)}{0.0091}X_5 - \underset{(1.71)}{0.004}X_6 - \underset{(1.49)}{0.0034}X_7 - \underset{(2.66)}{0.000088}X_9 - \underset{(0.21)}{2.95}X_{10}$

$\bar{R}^2 = 0.35$, SEE = 0.04, F = 3.90

(4') $Y_4 = -1{,}780.68 + \underset{(1.24)}{2.70}Y_2 - \underset{(2.94)}{4{,}791.11}Y_3 + \underset{(0.84)}{0.10}Y_5 - \underset{(1.80)}{1.43}X_1 + \underset{(1.23)}{0.32}X_3 + \underset{(1.54)}{66.31}X_5 - \underset{(0.37)}{1.60}X_6 - \underset{(1.02)}{4.62}X_7 + \underset{(0.38)}{19{,}830.68}X_{10}$

$\bar{R}^2 = 0.38$, SEE = 286.74, F = 4.29

(5') $Y_5 = 593.44 + \underset{(0.56)}{0.22}Y_1 + \underset{(0.01)}{17.86}Y_3 + \underset{(0.84)}{0.10}Y_4 - \underset{(2.79)}{0.58}X_1 + \underset{(0.92)}{0.022}X_4 - \underset{(1.39)}{18.46}X_5 - \underset{(1.31)}{3.25}X_6 - \underset{(0.03)}{0.067}X_7 - \underset{(1.20)}{15{,}034.51}X_{10}$

$\bar{R}^2 = 0.81$, SEE = 66.18, F = 23.90

Notes: Numbers in parentheses are t-ratios; $\bar{R}^2 = R^2$ adjusted for degrees of freedom; SEE = standard error of estimate; N = 49.

pupil expenditures, lower local revenues for education, lower rates of nonpublic enrollments, and lower bond sales. A surprising result is that school size is positively associated with the amount of state aid.

The only adverse effect of state aid that the data reveal is its impact on local incentives to raise revenue on a short- or long-term basis (Y_5 and Y_4, respectively). It appears to have a favorable effect on school size, expenditures, and public enrollments.

REFERENCES

Bishop, George A. (1964). "Stimulative versus Substitutive Effects of State School Aid in New England." *NATIONAL TAX JOURNAL 17* (June, 1964): 133-143.

Booms, Bernard H., and Hu, Teh-wei (1971). "Toward a Positive Theory of State and Local Public Expenditures: An Empirical Example." *PUBLIC FINANCE* 26:419-436.

Brazer, H. E. (1959). *CITY EXPENDITURE IN THE UNITED STATES.* Occasional Paper 66. New York: National Bureau of Economic Research.

Cohn, Elchanan (1968). "Economies of Scale in Iowa High School Operations." *JOURNAL OF HUMAN RESOURCES* 3 (Fall, 1968):422-434.

Cohn, Elchanan (1974). *ECONOMICS OF STATE AID TO EDUCATION.* Lexington, Mass.: D.C. Heath & Co. (in press).

Cohn, Elchanan, and Hu, Teh-wei (1973). "Economies of Scale, by Program, in Secondary Schools." *JOURNAL OF EDUCATIONAL ADMINISTRATION* 11 (October, 1973).

McMahon, Walter W. (1970). "An Economic Analysis of Major Determinants of Expenditures on Public Education." *REVIEW OF ECONOMICS AND STATISTICS* 52 (August, 1970) :242-252.

Miner, Jerry (1963). *SOCIAL AND ECONOMIC FACTORS IN SPENDING FOR PUBLIC EDUCATION.* Syracuse, New York: Syracuse University Press.

Renshaw, Edward F. (1960). "A Note on the Expenditure Effect of State Aid to Education." *JOURNAL OF POLITICAL ECONOMY* 68 (April, 1960): 170-174.

Riew, John (1966). "Economies of Scale in High School Operations." *REVIEW OF ECONOMICS AND STATISTICS* 48 (August, 1966): 280-287.

Sabulao, C. M., and Hickrod, G.A. (1971). "Optimum Size of School District Relative to Selected Costs." *JOURNAL OF EDUCATIONAL ADMINISTRATION* 9 (October, 1971): 178-191.

Sacks, Seymour (1972). *CITY SCHOOL, SUBURBAN SCHOOL: A HISTORY OF CONFLICT*. Syracuse, New York: Syracuse University Press.

SEX RATIO AT BIRTH IN RELATION TO FATHER'S OCCUPATION IN THE UNITED STATES: A SIMULATED ANALYSIS

M. E. El-Attar and S. M. El-Hakeem
Mississippi State University

Introduction

The union of the constant female X chromosome with either Y or X chromosome possessed by the male parent, which results in the first case with conception of a male and in the second case with conception of a female has established the belief that the male parent alone determines the sex of offspring (Eastman, 1963; Shettles, 1962). For a while it was believed that the fertilization of the ovum by an X or Y spermatozoon was random, with a 50 percent chance for either. But the excess of male births was noted by John Graunt about the turn of the seventeenth century (Bogue, 1969). Since that time, the topic has continued to attract the curiosity of outstanding investigators of demography as well as scientists in some of the closely related fields.

Statement of the Problem

The balance between male and female births has been shown to be almost stable from one calendar year to another, varying around 105 males to 100 females, with a little increase in births of males in war time compared to peace time (Tarver and Lee, 1968; Winston, 1931; Myers, 1947). Attempts to interpret the familiar occurrence of this ratio by any single factor have not been successful. Theories of divine intervention and natural compensation from the past have been refuted. Analyses of environmental factors and genetic factors have been attempted. The literature is rich in biological and sociological explanations. Biologists are trying to produce the desired sex in laboratory experiments or to detect the sex early in pregnancy (Edwards and Fowler, 1970; Edwards and Gardner, 1968; Rorvik and Shettles, 1971). Other biologists, together with demographers, sociologists, and psychologists, have taken advantage of vital statistics to relate such variables as age of mother, age of father, and birth order to the sex ratio of births. Findings of these studies could be summarized, in general terms, as follows:

1. An inverse relationship was observed between age of mother and sex ratio of births. In later writings the relationship was discovered to be spurious and was inferred differently by different writers (Novitski and Sandler, 1956; Novitski and Kimball, 1958; Myers, 1947; Tarver and Lee, 1968; Rubin, 1967).

2. In most studies an inverse relationship was observed between age of father and sex ratio of births. In some studies the inverse relation was found to be reversed when the father reached 40 years of age (Szilard, 1960; Novitski and Sandler, 1956; Novitski and Kimball, 1958; Novitski, 1953; Rubin, 1967).

3. Data on birth order, contrary to that for age of father, were easily available and were found to have an inverse relationship with sex ratio at birth. Correctly or not, the relation between mother's age and sex ratio at birth in many instances was reverted to birth order as it is directly correlated with mother's age (Tarver and Lee, 1968; Novitski and Sandler, 1956; Novitski and Kimball, 1958; Myers, 1954; Rubin, 1967).

4. A higher sex ratio at birth is regularly found for whites as compared with blacks, and is often attributed to the greater physical handicap of the socially disadvantaged black mother.

5. The sex ratio at birth rises with social class. This difference is ascribed to social factors rather than to innate differences. Among investigators arriving at these conclusions were Carberg and Lenhossek (cited in Winston, 1931).

6. Studies of data from different countries indicate a higher sex ratio in rural than in urban areas (Winston, 1931).

7. Sex ratio for legitimate births is higher than that for illegitimate births in both European and American data (Winston, 1931).

8. The idea of higher sex ratio in war time compared with peace time was not well established when comparison was made for different countries in different war periods (Panunzio, 1943; Myers, 1947).

Most of these studies were based on small samples, and, accordingly, conclusions were limited in scope. Moreover, since data on occupation of father in relation to the sex of his children have been very limited, there are no firm conclusions. The present study aims at this objective despite a lack of relevant data. In order to reach an approximate conclusion, a simulated analysis was undertaken. Such analysis must be considered as a suitable rather than a refined technique, since simulations are being described as crude structures (Fleisher, 1968). On the other hand, studies of the sex ratio of births are lacking in theory, and where theory is weak, simulations are believed to aid the investigator in the process of theorizing as well as in making "accurate, intelligent observations".

Construction of the Theoretical Model

The simulated analysis proposed in this study is based on a theoretical model which constitutes those variables to which the sex ratio of births is thought to be most closely related. According to this model the sex ratio at birth is conceptualized to be a result of one or any combination of two major sets of factors:

Endogenous Biological Factors

Endogenous biological factors in this study refer to innate factors, that is, elements pertaining to heredity and other inborn factors which exist in the parents from birth. These endogenous factors are considered to affect the sex of a child and include: (1) heredity,

(2) age of parents, and (3) child birth order.

Socio-economic and Exogenous Bio-psychic Factors

This set of factors constitutes the three component variables of socioeconomic status, namely occupation, education, and income. Only occupation of father is to be considered in the present analysis. The emergence of father's occupation as a prominent factor in this study stems from the fact that occupation is a factor which determines not only the socioeconomic position of an individual but his bio-psychic nature as well. These consequences characterize occupations "in urbanized societies in general, and in the United States in particular," where "occupations are differentiated on the basis of income, rights and privileges," and as the individuals's "image of himself as the holder of a particular specilized position in the division of labor" (Taylor, 1968; Becker and Carper, 1956). In other words, occupation of father may be assumed to have effects on the social and sexual life of parents so that the ability to conceive a male child or bring it to term is affected. A theoretical schematic model for this conceptualization is illustrated in Figure 1.

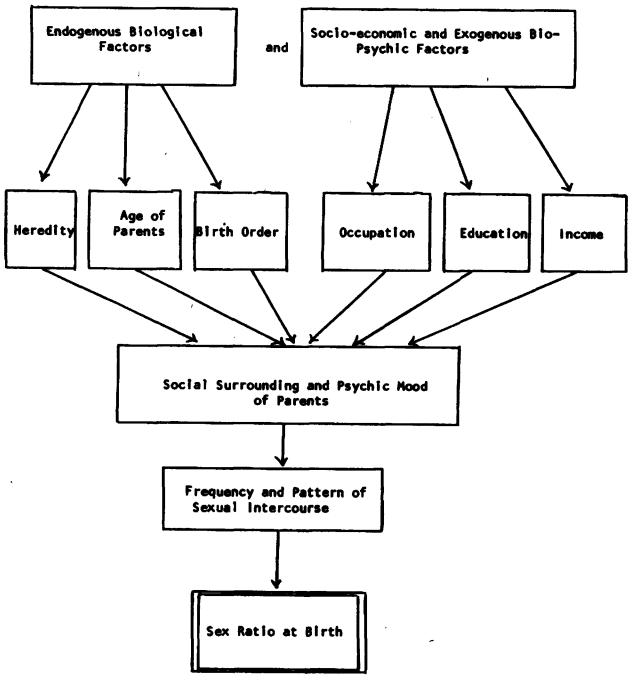


Figure 1-- Schematic Model for the Theoretical Conceptualization of the Relationship Between Sex Ratio at Birth and Selected Factors

The simulated analysis will use the following variables as a basic framework: (1) occupation of father, (2) age of parents, and (3) birth order.

Occupation

"An occupation is the social role(s) that constitutes a major focus in the life of adult members of society and that directly and/or indirectly" generates socioeconomic-and-biopsychic consequences to the performing actors (Hall, 1969). The biological consequences generated by occupations are termed exogeneous biological factors, to distinguish them from the endogenous (innate) biological consequences. Occupation as a concept indicates source of livelihood, career, status, prestige, and image. Obvious major differences exist among occupations. Most important to this study is the fact that some occupations require special mental ability while others stress manual dexterity. Certainly, the conditions of the job have an impact on the responses of the worker to his total life situation. Security and stability characterize more occupational groups; anxiety or insecurity are evident among others. The difficulty of dichotomizing occupational categories into manual and mental, and the absence of another suitable typology, were strong and convicingly good reasons for using the broad categories found in the census and available in the 1/1,000 sample (U.S. Bureau of the Census,1964). Sketches of the work conditions for each group are given in a recent study by the co-author (El-Hakeem, 1973).

Age

In most societies, age connotes biological and social elements. In the present analysis, the concept of "age" refers to the "biologic age", that is, "the person's relative functioning capacity as determined by the sum of genetic and environmental factors" (Petersen, 1969). The lack of accurate data on age of father has led to surveys with small samples and doubtful conclusions. However, the literature on age of father and sex ratio at birth is laden with conclusions which document inverse relation between age of father and that ratio. Regarding age of mother, no significant relation between sex ratio and mother's age was found. However, ages of both the father and mother are used as control variables in this study.

Birth Order

Birth order refers to "birth rank", that is, "higher order births or subsequent births are births occurring after the last specified order" (United Nations, 1958). In this study, only births of the current marriage are considered and used as a control variable.

The Hypotheses

It is observable in the American society that the father bound up in his job, anxious for success, and striving for prestige can be expected to have a lower frequency of intercourse than his fellow man, who is equal to him in every respect except that he has a manual job, and may be likely to spend more time with his family relaxing and seeking excitement and enjoyment far away from

his work. Even if the intellectual, ambitious father happens to have frequent sexual relations with his wife, one or both may be more likely to be under mental anxiety and stress, which affect their biophysical conditions and consequently may affect the sex ratio at birth.

In view of the above conceptualization as verbally stated and schematically outlined, the following hypotheses can be inferred.

Father's occupation has an effect on the sex of his child as follows: (a) husbands engaged in occupations which are characterized with extra job expectations tend to have low proportions of male offspring, (b) husbands engaged in routine work occupations tend to have high proportions of male offspring.

Source of Data, Universe of Analysis and Simulation

Given below are descriptions of source of data, universe of analysis, and the simulation process.

Source of Data

The source of data for this study is the 1/1,000 sample from the 1960 population census of the United States (U.S. Bureau of the Census, 1964). This sample was selected from a source file of the records of another sample (five percent) of population of the United States. The records in the source file were grouped by households in such a manner that a record for household head was followed by the records of all other members of the head's household. The sample includes 179,562 cases (individuals) grouped in households, including primary families, subfamilies, secondary families, and other (secondary individuals, non-inmates in group quarters, and inmates).

Univers of Analysis

The families selected for this study were limited to white families with children under 5 years of age living in urban areas, and who should satisfy the following criteria: (1) the father was married once; (2) the mother was married once; (3) both the father and mother were living and were counted in the sample; (4) the father was the chief income recipient and the head of the family; (5) the father had an occupation and an income; (6) the children represented in the family and counted at the time of the survey were equal to the number of children ever born by the mother.

The Simulation Process

Simulation is generally understood by people to be an "attempted replication of reality" (Goroff, 1973). It is a technique which enables the simulator to investigate some hypothesized relationship about phenomena abstracted from reality. "Simulation in research provides a controlled environment in which most parameters affecting the system may be examined and quantified (McCluskey, 1973). In the study of complex phenomena such as this one, where adequate data are not available, simulation as a tool permits one to explore the phenomenon and to specify the limitations of the existing data through the "imitation of a real process". (Dutton and Briggs, 1971; Sheps,1971).

The simulation process in the present study has started out with two steps that constituted the organization of the data. The first step involves selection of the desired data from an initial population of households, stored on five magnetic tapes. From these five tapes, a seven-track magnetic tape was built on Univac 1106. The tape includes all children in the 1/1,000 sample who belong to parents where marriage to each other is for the first time, with no infant mortality, with father having an occupation and being chief income recipeint. Children selected according to the above criteria included 33,081 children belonging to 13,533 families.

The tape is built in such a manner that each child has one record containing the following information as provided in the 1/1,000 documentation (U. S. Bureau of the Census, 1964): Household number, family number, size of place of residence, father's year of marriage, mother's year of marriage, father's marital status, mother's marital status, father's age, mother's age, father's race, mother's race, father's education, father's occupation, father's income, child's quarter of birth child's sex, child's age, children ever born by the mother of the child, and child's birth order, which is arranged according to each child's age together with his siblings.

The computer program written for processing all operations in this step is available from the authors on request. The logical objectives of the program are: (1) To check on every record of the total sample; (2) When the computer read a record of a head of family or wife of a head, eligibility for the above criteria was tested, data were stored, and a search was then started for the partner in order to test eligibility and to store data required about him or her; (3) When the computer detected any ineligible record in item "b", it was required to drop all of the records for the family, to cancel any stored data, and to start search for the next family; (4) Since some of the children were not arranged on the original tapes according to age, records were checked and reorganization of children by age was made on the new tape; (5) For every family selected, the number of children counted was checked against the number of children ever born to the mother. Whenever these two items were not equal, the family was dropped from the analysis.

From the major information provided above, eleven variables were taken to form a basis to the second step. The eleven variables: type of residence, mother's age, mother's race, father's age,

father's race, father's education, father's occupation, father's income, sex of child, age of child, and birth order of each child.

In the second step the simulation task is to specify the type and characteristics of children to be chosen for the analysis. To enhance reliability of the data, it was proposed to limit the sample to white children under five years of age in 1960, who were residing in urban areas. The sample includes 7,062 children (21.3 percent of the sample selected in the first step) belonging to 4,463 families (33 percent). Before the tabulations were done, both age of mother and age of father were computed as of the time of the child's birth by subtracting the child's age in 1960 from the mother's and father's age in the same year. The data on number of families and children and proportions that are males cross-classified by occupation and age of father, age of mother, and/or birth order are given in Tables 1,2, 3, and 4. Commentary on these tables is given in analysis of findings below.

TABLE 1. NUMBER OF SELECTED FAMILIES AND CHILDREN UNDER 5 YEARS OF AGE AND THE PROPORTION OF MALE CHILDREN BY OCCUPATION OF FATHER IN URBAN AREAS OF THE UNITED STATES, 1960[a]

Occupation of Father	Families		Children		
	Number	Percent	Number	Percent	Male Proportion
Professional, technical & kindred workers	708	15.86	1,113	15.76	.4978
Farmers & Farm Managers	132	2.96	213	3.02	.5070
Managers, Officials, and Proprietors, except farm	509	11.40	769	10.89	.5137
Clerical & kindred workers	318	7.13	506	7.17	.4881
Sales workers	345	7.73	532	7.53	.4944
Craftsmen, Foremen, & kindred workers	1,109	24.85	1,728	24.47	.5098
Operatives & kindred workers	938	21.02	1,511	21.40	.5175
Service workers, except private household	160	3.58	261	3.70	.4943
Laborers, including farm	244	5.47	429	6.07	.5128
Total	4,463	100.00	7,062	100.01	.5068

TABLE 2. NUMBER OF CHILDREN UNDER 5 YEARS OF AGE AND PROPORTION OF MALES AMONG THEM BY OCCUPATION AND AGE OF FATHER IN URBAN AREAS OF THE UNITED STATES, 1960[a]

Occupation of Father	Age of Father									
	20 or less		21-30		31-40		41 & Over		Total	
	No.	M.P.[b]	No.	M.P.	No.	M.P.	No.	M.P.	No.	M.P.
Professional wkrs.	11	.5455	561	.5116	485	.4907	56	.4107	1,113	.4978
Farmers & farm mgrs.	8	.7500	100	.5300	85	.4824	20	.4000	213	.5070
Managers, etc.	10	.6000	362	.5028	337	.5134	60	.5667	769	.5137
Clerical wkrs.	14	.2857	317	.4700	153	.5556	22	.4091	506	.4881
Sales wkrs.	19	.5789	308	.5065	182	.4615	23	.5217	532	.4944
Craftsmen, etc.	57	.5614	975	.5046	613	.5057	83	.5663	1,728	.5098
Operatives, etc.	105	.5429	878	.5159	453	.4989	75	.6133	1,511	.5175
Service wkrs., except pvt. househd.	9	.3333	158	.4557	83	.5663	11	.6364	261	.4943
Laborers, including farm	26	.5385	247	.5182	133	.4962	23	.5217	429	.5128
Total	259	.5367	3,906	.5049	2,524	.5032	373	.5308	7,062	.5068

TABLE 3. NUMBER OF CHILDREN UNDER 5 YEARS OF AGE AND PROPORTION OF MALES AMONG THEM BY OCCUPATION OF FATHER AND AGE OF MOTHER IN URBAN AREAS OF THE UNITED STATES, 1960[a]

Occupation of Father	Age of Mother									
	20 or less		21-30		31-40		41 & Over		Total	
	No.	M.P.[b]	No.	M.P	No.	M.P.	No.	M.P.	No.	M.P.
Professional wkrs.	58	.5000	727	.5021	318	.4937	10	.3000	1,113	.4978
Farmers & farm mgrs.	23	.6087	127	.4803	60	.5167	3	.6667	213	.5070
Managers, etc.	45	.4667	454	5022	252	.5317	18	.6667	769	.5137
Clerical wkrs.	61	.4590	336	.4847	112	.5268	7	.2857	506	.4881
Sales wkrs.	47	.4894	352	.4943	130	.4923	3	.6667	532	.4944
Craftsmen, etc.	220	.5545	1,049	.4881	446	.5471	13	.2308	1,728	.5098
Operatives, etc.	273	.5092	894	.5212	323	.5139	21	.5238	1,511	.5175
Service wkrs., except pvt. househd.	37	.4324	174	.4713	46	.6522	4	.2500	261	.4943
Laborers, including farm	82	.5122	256	.5273	85	.4706	6	.5000	429	.5128
Total	846	.5130	4,359	.5003	1,772	.5220	85	.4588	7,062	.5068

TABLE 4. NUMBER OF CHILDREN UNDER 5 YEARS OF AGE AND PROPORTION OF MALE AMONG THEM BY OCCUPATION OF FATHER AND BIRTH ORDER IN URBAN AREAS OF THE UNITED STATES, 1960[a]

Occupation of Father	Birth Order											
	1		2		3		4		5		6+	
	No.	M.P.[b]	No.	M.P.	No.	M.P.	No.	M.P.	No.	M.P.	No.	M.P.
Professional wkrs.	289	.5017	376	.5266	257	.4630	122	.5082	39	.4103	30	.4667
Farmers & farm mgrs.	35	.4857	61	.5738	56	.4821	28	.3929	16	.7500	17	.3529
Managers, etc.	137	.4818	268	.5299	213	.5023	85	.6000	38	.5526	28	.2857
Clerical wkrs.	144	.4792	184	.4837	115	.4261	41	.6098	19	.6842	3	.6667
Sales wkrs.	115	.5043	192	.4688	115	.5304	67	.5373	25	.4000	18	.4444
Craftsmen, etc.	371	.5067	553	.5280	427	.4988	224	.5045	83	.4819	70	.5000
Operatives, etc.	370	.5324	486	.4897	325	.5292	177	.5311	82	.5488	71	.5070
Service wkrs., except pvt. househd.	49	.4898	86	.4651	70	.5857	29	.5172	19	.3158	8	.3750
Laborers, including farm	99	.5252	131	.5115	87	.5402	45	.4444	28	.6429	39	.4103
Total	1,609	.5071	2,337	.5096	1,665	.5021	818	.5220	349	.5186	284	.4507

Analysis of Findings

The findings of the study and tests of the hypotheses are given in two parts: The first is a descriptive interpretation of the data; the second is statistical analysis couched in a correlation method.

I. Descriptive Interpretation

Tables 1, 2, 3 and 4 are the main data source for this part of the analysis. The tables give the proportion of male children cross-classified by occupation and age of father, age of mother, and birth order.

Male Proportion and Occupation of Father. -- Table 1 gives the number of selected families and children under 5 years of age and the proportion of male children by occupation of fathers in urban areas of the United States in 1960. The two major occupational groups of farm laborers and laborers are lumped together. The proportion of male children is .507, that is, for every 1,000 children there are 507 males and 493 females. The comparative standing of the proportion of male children in an occupation, compared with the male proportion of children for all occupations, is of assistance in evaluating the level of differentiation among the different occupations of fathers as regards the male proportion of their children. Table 1 shows that among non-manual workers, clerical workers have the lowest male proportion among their children, followed by sales workers and professional, technical and kindred workers. Managers, officials and proprietors are the only non-manual group in which fathers have more boys than girls among their children. Fathers occupied in service occupations are the only group among the manual workers who have low proportion of male children.

Proportion of Males by Age and Occupation of Fathers. -- Table 2 provides the number of children under 5 years of age and the proportion of males among them by occupation of fathers in urban areas of the United States in 1960. The proportion of male children for the total of each age group shows a gradient from the high of .5032 for fathers who are 31 to 40 years of age to .5367 for fathers who are 20 years old or less.

Because of the differences in occupational tasks, even in the same occupation, and of differential

effects of stressful norms (norms emphasizing competition and achievement of success) on different individuals (see Brown, 1954 who indicated that some individuals are temperamentally more liable to stress than others) a sharper gradient among the occupation groups with regard to male proportion among children under 5 years of age is evident. These results are in accord with the conceptual framework formulated above for this study, provide support for it, and demonstrate its relevance to this research.

Proportion of Males by Age of Mother and Occupation of Father. Table 3 gives the number of children under 5 years of age and the proportion of males among them classified by age of mothers and occupation of father in urban areas of the United States in 1960. The proportion of male children for the total of each age group shows a gradient ranging from the low .4588 for mothers who are 41 years old and over to the high .5220 for mothers who are 31 to 40 years of age. Comparing the proportion of male children by age of mother with that by age of father, one finds that in both cases an inverse relationship exists between male proportion and age (except in the age group 31 to 40 for mothers and 41 and over for fathers). The age group 31 to 40 is the age at which females married to men engaged in different occupations achieve the highest male proportion (.5222) among their children, whereas a direct relationships between the proportion of male children and age of father is revealed by the age group 41 and over. Generally stated, the proportion of male children at age 41 and over is inversely related to age in the case of mothers and directly related in the case of fathers.

With regard to the male proportion in the individual occupational categories by age of mother, it is worthy to indicate that the generated pattern does not follow that of the total. Hence, it could be said that father's occupation might be a basic factor in producing such differentiated pattern in the proportion of male children among the different occupations.

Male Proportion by Parity and Father's Occupation. -- This section is concerned with the differences in the proportion of male children as associated with occupation of fathers and birth orders. Table 4 gives the number of children under five years of age and the proportion of males among them cross-classified by occupation of father and birth order in urban areas of the United States in 1960. The table shows that the proportion of male children according to birth order (except the third and sixth and above orders), when occupation effect is randomized, exceeds the proportion of male for all occupations (.5068 taken as an average). The fourth order is the modal value for the parity distribution of male proportions. Moreover, the inverse relation between the proportion of male children and birth order, as claimed by the literature, does not hold here, as male proportion is higher in the second order than in the first and third order, and the fourth order is higher than any order in the analysis. The antithesis of an inverse association is supported also by the male proportion in the fifth order which is higher than the first, second, and third orders. This disagreement with the literature is in accord with the conceptual framework of this study which proposes the proportion of male children to be the product of a combination of biological and social factors.

Regarding the profile of the proportion of male children by parity and father's occupation, the data in Table 4 reveal that the above pattern still holds for the occupational categories. However, the influence of occupation of fathers is noticeable in the differentiated magnitudes of the proportions of male children within each birth order.

II. Statistical Analysis

In preparation for a statistical analysis, the data were simulated in a factorial design as illustrated by Table 5. A five-dimension array was prepared

TABLE 5. ILLUSTRATION OF COMPUTER SIMULATED DATA WHERE PROPORTION OF MALE CHILDREN UNDER 5 YEARS OF AGE IS A CRITERION VARIABLE, AND AGE OF MOTHER, AGE OF FATHER, AND BIRTH ORDER ARE PREDICTORS

Occupation of Father	Simulated Sample Size	Age of Mother	Age of Father	Education of Father x_3	Income of Father x_4	Birth Order	Number of Sons	Number of Daughters	Male Proportion
Professional wkrs.	355								
Farmers and farm mgrs.	158								
Managers, etc.	341								
Clerical wkrs.	208								
Sales wkrs.	265								
Craftsmen, etc.	495								
Operatives, etc.	484								
Service wkrs. except pvt. household.	168								
Farm laborers	61								
Laborers	220	In each occupational group, different ages, starting with age 15 (mid-point) and spaced by 10-year interval, were associated with possible combination of father's age.	In each occupational group, different ages, starting with age 15 (mid-point) and spaced by 10-year interval, were associated with every possible age of mother.	Number of years of school completed by every possible father in each occupation group.	Common logarithms of father's income as reported in the census in each occupation group.	Birth order in each possible combination of father and mother in each occupation group.	Number of male children under 5 years of age in each possible combination in each occupation group.	Number of female children under 5 years of age in each possible combination in each occupation group.	Number of male children under 5 years of age divided by total number of children in each corresponding family combination in each occupation group.

containing proportion of male children under five years of age as a criterion variable and five predictors, namely, mother's age, father's age, father's education, father's income, and birth order. It is worthy at this point to remember that occupation of father was used as a replicated variable. The proportion of male children belonging to one combination of independent variables was considered as one observation. Grouping the data in this way resulted in 2,755 observations, distributed as follows: professional, technical and related workers, 355; farmers and farm managers, 158; managers, officials and proprietors, 341; clerical and kindred workers, 495; operatives and kindred workers, 484; service workers, 168; farm laborers, 61; laborers, except farm and mine, 220.

The relation between the male proportion and three of the five independent variables, namely, age of mother, age of father, and birth order is assumed to be a functional linear relationship.

The relationships between proportion of male children under 5 years of age and the three selected independent variables in each major occupation group are given in Table 6. Differentiation between occupations is reflected in the magnitude and signs of the coefficients of simple and partial correlations. And although the association is not significant in most occupations the hypothesis that father's occupa-

TABLE 6. SIMPLE AND PARTIAL CORRELATION COEFFICIENTS BETWEEN PROPORTION OF WHITE MALE CHILDREN UNDER 5 YEARS OF AGE AND AGE OF MOTHER, AGE OF FATHER, AND PARITY FOR EACH OCCUPATION GROUP IN URBAN AREAS OF THE UNITED STATES, 1960

Occupation of Father	Correlation Coefficients					
	Age of Mother		Age of Father		Birth Order	
	Simple	Partial	Simple	Partial	Simple	Parital
Professionals, etc.	-.081*	-.068	-.043	.077	-.027	.007
Farmers and farm mgrs.	-.031	.073	-.126*	-.148**	-.019	.025
Managers, etc.	.040	.041	.027	.012	-.029	-.050
Clerical wkrs.	.011	-.004	.006	-.014	.067	.070
Sales wkrs.	-.060	.070	-.003	-.056	.042	.015
Craftsmen, etc.	-.028	-.029	-.009	.010	-.007	.003
Operatives, etc.	.037	-.007	.066*	.052	.034	.014
Service wkrs., except pvt. houshd.	.112*	.075	.160+	.137**	-.077	-.145**
Farm laborers	-.089	-.051	-.080	-.034	-.050	.010
Laborers	-.002	.000	.001	.008	-.021	-.023
Total	-.027*	.012	-.052x	-.041+	-.038+	-.023*

*Significant at 25 percent level.
**Significant at 10 percent level.
+Significant at 5 percent level.
xSignificant at 1 percent level.

tion has a different influence on the proportion of male children among his offsprings is partially confirmed. For example, the signs of the partial correlation coefficients (which provide the direction of the relationship) are in opposition to each other in most of the occupational groups. This indicates that, with other things being equal, the influence of a given occupation--for example, professional, technical and kindred workers--on the sex of the child when the age of mother is considered will be different from that of a certain other occupation--for example, service workers. Confirmation of this finding is readily conspicuous from the difference in magnitudes and signs of the simple and partial correlation coefficients obtained from the total and those obtained from the individual occupations.

Summary and Conclusion

In this paper, a theoretical model which embraced those variables to which the sex ratio at birth might be most closely related was developed, and relationships were examined between proportion of male children and the stated variables. Because of the belief that the 1/1,000 sample might be an unreliable source of data for the presetn study, a simulated analysis was used. Both the descriptive and statistical analyses partially supported the stated hypotheses.

Among the most severe drawbacks of census data used in this study were the following: (1) Occupation of the father pertained to census time rather than the time of conception; (2) The data did not indicate whether the father and mother were working at time of conception, and if they were, what the occupational group was for each of them; (3) Occupation as given in the census group is a nominal variable and it is deemed necessary to develop an occupational score according to which occupations can be differentiated, stratified, or ranked with regard to the stress they exert on their incumbents.

More clear-cut findings of the influences of father's occupation on the sex of his children at birth at various ages of both parents as well as parity, would be made possible with the aid of more adequate data. Such data can be established by means of a special survey or, for a lower quality of data, from a special tabulation of vital statistics records.

More adequate data may enable us to construct a general statement according to which the probability of giving birth to a male can be determined by age of father, age of mother, parity, and socioeconomic status of the father.

Acknowledgements

The authors are indebted to Professors E.S. Lee and J. D. Tarver for providing the data and for their instructive comments. The helpful comments of H. Armstrong are acknowledged. Any errors of fact or interpretation are our responsibility.

Footnotes

[a]Source: Compiled and computed from U.S. Bureau. U.S. Census of Population: 1960. 1/1,000, 1/10,000, Two National Samples of the Population of the United States. Washington, D. C., 1964.

[b]M.P. stands for male proportion.

[c]Variables included in the simulation process for the purpose of formulating analytical observations but their influence on the dependent variable was not considered in the present analysis.

References

Becker, Haward S. and James W. Carper. 1956. "The Development of Identification With An Occupation," American Journal of Sociology, 61 (January), 289-298.

Blau, Peter M. and Otis Dudley Duncan. 1967. The American Occupational Structure, New York: John Wiley and Sons, Inc.

Bogue, Donald J. 1969. Principles of Demography, New York: John Wiley and Sons, Inc.

Brown, J.A.C. 1954. The Social Psychology of Industry, Baltimore: Penguin Book.

Cohen, M. R. 1964. "Detection of Ovulation by Means of Cernical Mucous and Basal Body Temperature," Ovulation, 26, 291-298.

Dutton, John M. and Warren G. Briggs. 1971. "Simulation Model Construction," pp. 103-126 in John M. Dutton and William H. Starbuck (ed.),

Computer Simulation of Human Behavior, New York John Wiley and Sons, Inc.

Eastman, Nicholson J. 1963. Expectant Motherhood, Boston: Little, Brown and Company, pp. 18-22.

Edwards, Robert and Richard Gardner. 1968. "Choosing Sex Before Birth,"New Scientist, 2 (May) 218-222.

Edwards, R.G. and Ruth E. Fowler. 1970. "Human Embroys in the Laboratory," Scientific American (December): 45-54.

El Hakeem, S.M. 1973. "Sex Ratio at Birth by Occupation of Father: A Test of the Usefulness of the 1/1,000 Sample." Unpublished M.A. Thesis, University of Georgia.

Fleisher, Aaron. 1968. "The Uses of Simulation," pp. 187-189 in James M. Beshers (ed.), Computer Methods in the Analysis of Large Scale Social Systems, Cambridge: The M.I.T. Press.

Goroff, Norman. 1973. "Simulated Incarceration Experiences," Simulation and Games, 4 (March): 59-70.

Hall, Richard H. 1969. Occupations and the Social Structures, Englewood Cliffs: Prentice-Hall, Inc.

Shettles, B. Landrum. 1962. "Human Spermatozoa Population," Internationa Journal of Fertility, 7 (April-June): 175-187.

Greenwood, Ernest. 1962. "Attributes of a Profession," pp. 219-224 in Sigmund Nosow and William H. Form (eds.), Man, Work, and Society, New York: Basic Books.

Lorimer, Frank. 1959. "The Development of Demography," In Philip M. Hauser and Otis Dudley Duncan (eds.), The Study of Population, An Inventory and Appraisal, Chicago: The University of Chicago Press, pp. 129-130.

McCluskey, Michael R. 1973. "Perspectives on Simulation and Miniaturization," Simulation and Games, 4 (March): 19-36.

Masters, W.O. 1967. "Note in Medical Section," Time, (April 7): 84.

Myers, R. J. 1947. "Effect of the War on the Sex Ratio at Birth," American Sociological Review, 12 (February): 40-43.

Myers, R.J. 1954. "The Effect of Age of Mother and Birth Order on Sex Ratio at Birth," The Milbank Memorial Fund Quarterly, 32 (July): 275-281.

Novitski, E. 1953. "The Dependence of the Secondary Sex Ratio in Humans on the Age of Father," Science, 117 (May 15): 531-533.

Novitski, E. and L. Sandler. 1956. "The Relationship Between Parental Age, Birth Order, and the Secondary Sex Ratio in Human." Annals of Human Genetics, 21: 123-31.

Novitski, E. and A. W. Kimball. 1958. "Birth Order, Parental Ages, and Sex of Offspring," American Journal of Human Genetics, 10: 268-275.

Panunzio, Constantine. 1943. "Are More Males Born in Wartime?" The Milbank Memorial Fund Quarterly, 21 (July): 281-291.

Parkes, A.S. 1963. "The Sex Ratio in Human Population," Man and His Future, 158.

Petersen, William. 1969. Population, New York: The Macmillan Company.

Pitts, F. N. 1969. "The Biochemistry of Anxiety," Scientific American, 220 (February): 69-75.

Rock, J. and D. Robinson. 1965. "Effect of Induced Intrascrotal Hyperthermia on Testicular Function in Man," American Journal of Obstetrics and Gynecology, 93 (November): 793-801.

Rorvik, David M. and Landrum B. Shettles. 1971. Your Baby's Sex: Now You Can Choose, New York: Dodd, Mead and Company.

Salz, Arthur. 1962. "Occupations in Their Historical Perspective," pp. 58-63, in Sigmund Nosow and William H. Form (eds.), Man, Work, and Society, New York: Basic Books.

Schuster, D. H. and Locky Schuster, 1969a. "Study of Stress and Sex Ratio in Humans," Proceedings, 77th Annual Convention, American Psychological Association, 335-336.

Schuster, D. H. and Locky Schuster. 1969b. "Theory of Stress and Sex Ratio," Proceedings, 77th Annual Convention, American Psychological Association, 223-224.

Sheps, Mindel C. 1955. "An Examination of Some Methods of Comparing Several Rates or Proportions," Biometrics, 15 (March): 87-97.

Sheps, Mindel C. 1971. "Simulation Methods and the Use of Models in Fertility Analysis," International Population Conference, The International Union for the Scientific Study of the Population, London, 1969, 1: 53-64.

Selye, H. 1956. The Stress of Life, New York: McGraw-Hill.

Smith, T. Lynn. 1948. Population Analysis, New York: McGraw-Hill Book Company, Inc.

Snyder, R. G. 1961. "The Sex Ratio of Offspring of Pilots of High Performance Aircraft," Human Biology, 33: 1-10.

Spiegelman, Mortimer. 1955. Introduction to Demography, Chicago, Illinois: The Scoiety of Actuaries.

Szilard, Leo. 1960. "Dependence of the Sex Ratio at Birth on the Age of Father," Nature, 186 (May 21): 649-650.

Tarver, James D. and Che-Fu-Lee. 1968. "Sex Ratio of Registered Live Births in the United States, 1942-63," Demography, 5 (1): 374-381.

Taylor, M.A. 1969. "Sex Ratio of New Borns: Associated With Prepartum and Postpartum Schizophrenia," Science, 164 (May 9): 723-724.

U.S. Bureau of the Census. 1964. U.S. Census of Poppulation: 1960. 1/1,000, 1/10,000, Two National Samples of the Population of the United States, Description and Technical Documentation, Washington, D. C.: U.S. Government Printing Office.

U.S. Bureau of the Census. 1963. U.S. Census of Population: 1960, Detailed Characteristics, United States Summary, Final Report PC(1)-1D, Washington, D.C.: U.S. Government Printing Office.

United Nations. 1958. Multilingual Demographic Dictionary, English Section, New York: United Nations Publication, Sales N.: 58. XIII. 4.

Winston, Sanford. 1931. "The Influence of Social Factors Upon the Sex Ratio at Birth," The American Journal of Sociology, 7 (July): 1-21.

SAMPLING A RARE POPULATION: A GENERAL STRATEGY ILLUSTRATED BY A CASE STUDY

Eugene P. Ericksen, Institute for Survey Research, Temple University

1. Introduction

Studies concentrating on or restricted to subgroups of a total population comprise an increasing share of sample survey activity. At the Institute for Survey Research, we are seldom asked to carry out a straightforward cross-sectional study of the national household population where each household would have equal probability of selection. Instead, our national studies have focused on, for example, demographic subgroups defined by age, race, and sex, homeowners, or physicians concentrated in certain fields.

In such restricted studies, it is usually necessary to screen certain households based on information obtained from a door-answerer. This screening process is time-consuming, and if an extensive amount of screening is needed, the process can significantly increase the average cost of obtaining one interview. However, if the target population is at least partially segregated, the areas where it is concentrated can be sampled at a higher rate, reducing the amount of necessary screening and reducing the increase in the average cost per interview. The strategy is to form separate strata depending on variations in concentration of the target population and then to apply the strategy of optimal allocation (Hansen, Hurwitz, and Madow, 1953, section 6E; Cochran, 1963, pp. 95-97; Kish, 1964, sections 3.5, 3.6). Optimal allocation will bring about gains if the areas of concentration can be identified, and if costs vary directly with the amounts of concentration. However, it is important that a large proportion of the target population live in such segregated areas. Kish (1964, pp. 409-410) provides an illustration where the target population was high-income households. It was easy to identify the most exclusive neighborhoods, but most of the rich were scattered elsewhere and the gains from oversampling the exclusive neighborhoods were small.

Cochran (1963, p. 95) shows that in stratified sampling the variance of the overall mean of a given study variable is minimized when the specified sampling rate in a given stratum h is:

$f_h = n_h/N_h$ is proportional to $S_h/\sqrt{c_h}$, (1.1)

where

n_h = the expected number of selected observations in stratum h,

N_h = the total population in stratum h,

S_h^2 = the variance of the variable in stratum h,

c_h = the average cost of one interview in stratum h.

In most practical sampling situations, the between-stratum variations in S_h are minimal, so the sampling rates are usually functions of the cost variations. In computing these costs, it is important to include the costs of all components which vary with the numbers of interviews collected including the listing of households, travel, and the costs of coding.

2. The Specific Sampling Problem

The specific sampling problem providing our illustration was the construction of a sample of 15 through 19-year-old females living in households. The specified ratio of white to black respondents was two to one. The subject matter of the study, fertility practices and expectations, has been reported in part by Zelnik and Kantner (1972).

In order to obtain white and black interviews at the specified ratio, we estimated that it would be necessary to oversample blacks at 3.86 times the rate at which white respondents were to be selected. We also estimated that an eligible respondent would be found in one household in seven, and that completion rates would be 75 per cent for respondents of both races. Based on these computations, it was estimated that it would be necessary to visit $7/(.885 \times .75) = 10.55$ households to obtain one white interview and $7/(.115 \times .75) = 81.16$ households to obtain one black interview.

The increase in the number of required households to be contacted for one black interview was due to the necessity of screening out households with otherwise eligible white respondents. This screening would have been concentrated in predominantly white areas with few potential black respondents, and would have increased costs by an inordinate amount. It was therefore decided to subdivide the household population into two strata, with Stratum 1 to include predominantly white areas where blacks and whites were to be sampled at the same rate, and with Stratum 2 to include those areas where blacks lived in sufficient concentrations where they could be oversampled without undue increases in the amount of necessary screening. Whites in Stratum 2 were to be selected at the same rate as in Stratum 1, but blacks were to be selected at a rate 3.86 times greater.

It was then necessary to estimate the amounts of screening and increases in costs which would be obtained in areas varying in the proportions of households which were black. Because white respondents were to be selected at a constant, lower rate throughout, the additional costs of screening by race were all applied to the black interviews. The following assumptions were made:

a. Each cluster of households, or "listing area," would include an expected 100 households yielding an expected 9 or 10 interviews if there were no screening on

the basis of race.

b. Three hours would be required to list all households in a given listing area, 15 minutes would be required to determine for each household whether or not an eligible respondent lived there who was willing to be interviewed, and a total of 2 hours would be required to complete and code the interview.

c. The costs per hour would be the same for the various components of the total process.

The time necessary to complete each black interview was computed for listing areas varying in racial composition. These computations are shown in Table 1. There we can see that the expected time necessary to obtain a black interview rose sharply when the proportion of households which were black fell below 25 per cent. Applying equation (1.1), when 3 per cent of all households were black, the optimal sampling rate was 1/3.83 times the rate at which blacks would be sampled in a totally black area. If the proportion of black households was 10 per cent or less, the optimal sampling rate for blacks was closer to the sampling rate for whites than it was to the specified oversampling rate for blacks which was 3.86 times greater. It was therefore decided that if it were reasonably clear that a given listing area included fewer than 10 per cent black respondents, it would be placed in Stratum 1, and would otherwise be placed in Stratum 2.

The actual stratification procedure was composed of two stages. Estimates of the racial composition for the selected primary sampling units (psus) included in the national sampling frame were computed on the basis of (1) the 1960 Census, which was unfortunately 10 years out-of-date at the time, and (2) estimates of the racial composition of other listing areas in the same psus which had been used in recent surveys. If we confidently estimated, given the limitations of the data, that a psu was less than 3 per cent black, that psu was placed in Stratum 1, and listing areas were selected at 1/3.86 times the Stratum 2 rate. Household listings were then obtained from all selected listing areas in the two strata, and estimates of the racial composition were obtained. A certain amount of error in these estimates was expected, but if it was estimated that a given listing area in Stratum 2 included fewer than 10 per cent black households, and if this estimate did not contradict estimates computed from census data and past surveys, the listing area was transferred to Stratum 1. Listing areas thus transferred to Stratum 1 were then subselected at the rate of 1 in 3.86. Within Stratum 2, white households were subselected at this rate at the time of interviewing. Therefore, the difference in the sampling procedure for whites in the two strata was that in Stratum 1, all subselection was done in advance and no screening of households was necessary, whereas in Stratum 2, the sampling rate for households was 3.86 times greater, and households including potential white respondents were subselected in the field at the rate of 1/3.86.

3. Results: Numbers of Interviews

Two thousand nine hundred and fifteen (2,915) interviews were obtained with white respondents and 1,438 interviews with black respondents. Two thousand five hundred and thirty-five (2,535) interviews were obtained in Stratum 1, of which 58, or 2.29 per cent, were with black respondents. Multiplying the 438 white respondents in Stratum 2 and the 58 black respondents in Stratum 1 by 3.86, we estimate that 45 per cent of potentially eligible respondents in Stratum 2 were black, and that 86 per cent of all potentially eligible black respondents lived in Stratum 2.

The classification of areas into the two strata thus appears to have been accurate. Multiplying the 1,438 black interviews by the estimated 81.16 households necessary to obtain one interview, it would have been necessary to include 116,708 households in the sample had we not followed the stratification procedure. Had all blacks been selected at the higher rate, $1,380 + 58 \times 3.86 = 1,603.88$ interviews with blacks would have been obtained, but the sample would have had to include 130,170 households. Ninety-two thousand five hundred and ninety-nine (92,599) housing units were in fact listed. Of these, 11,594 were eliminated from the sample on the basis of the enumerator's estimate of the racial composition of listing areas, leaving a final sample of 81,005 housing units, 56.33 per black interview.

After the survey was completed, the racial composition of all listing areas included in Stratum 2 was calculated from screening forms filled out for each listed household, and compared to the estimates made by enumerators when listing households. These estimates tended to be accurate, with the mean absolute error being 13.39 per cent, the mean actual error (actual minus estimated percentage black) 3.08 per cent, and the median absolute error 6.03 per cent. Of the 477 estimates, 105 had greater than 20 per cent error, 20 had greater than 50 per cent error, and 7 had larger than 80 per cent error. Many of the larger errors contradicted the estimate expected on the basis of census and past survey data, and where there was doubt, a listing area was retained in Stratum 2. In most such cases, the error turned out to be a case of the interviewer giving the percentage black where the percentage white was intended.

The relative effects of actual racial composition of listing area, race of interviewer, region of the U.S., and central city-suburban status on the accuracy of the estimates of racial composition were measured by using these as explanatory variables in a multiple regression equation used to estimate the size of the absolute errors. This exercise was limited to SMSAs, because black interviewers did not work in non-metropolitan areas. The results are presented in Table 2. There we see that only one variable,

the actual racial composition of an area, provided any substantial clue to the accuracy of the estimate. The closer the actual composition of the area was to 50 per cent, the greater was the chance of error, and the other characteristics added only negligible amounts of explanation. When the larger, presumably random errors were eliminated from consideration, the explanatory power of the variables increased noticeably.

4. Results: Costs Per Interview

A computer record was kept of the results of calls of each listed address. Records of the salaries and expenses paid to each interviewer were also kept for both the household listing and interviewing phases of the study. The results of calls and salaries and expenses were aggregated to the psu level for computation of costs per interview. Where an interviewer worked in more than one psu, the psus were combined, and the full sample of 126 psus was thus reduced to 115 units.

It was estimated that the average cost of coding one interview was $4. Adding this cost to the tabulated costs of listing and interviewing, the average cost per interview of these components was $27.27. In the 69 psus where all listing areas were included in Stratum 1, the average cost per interview was $22.49, and in the remaining areas the average cost was $31.67. Assuming an average cost per white interview of $22.49 in these remaining 46 areas, the average cost per black interview rose to $36.21.

Regression equations estimating the average costs per interview are presented in Table 3. There we see that the increases in costs due to screening and to nonresponses were comparable, and that small reductions in costs were obtained in psus where the number of interviews obtained was large. Because the variation in the amount of screening was much greater than the variation in the other two variables, this was the primary determinant of variations in cost per interview.

We are now in a position to evaluate our decision regarding the optimal cutting point for stratifying listing areas. Within Stratum 1, an average of 10.5 households were screened out on the basis of age for every interview collected and an average of .67 nonresponses, usually refusals or cases where the presence of an eligible respondent could not be determined, were obtained per interview. Applying equation 5 from Table 3, the average per interview cost in a given psu in the absence of screening on the basis of race could then be expected to be

$$\hat{Y} = 11.79 + (1.18)\ (10.5) + (1.38)\ (.67) - .03K$$
$$= 25.09 - .03K \text{ dollars, where}$$

K = the number of interviews obtained in the psu.

Within Stratum 2, the presence of white households increased the amount of screening in proportion to the number of such households. The formula expressing the expected increase is

$$I = W + (I-W)/3.86 \text{ where} \qquad (4.1)$$

W = the proportion of households which are black, and

10.5/1 = the expected number of screened households per interview.

For example, in totally white areas, I = .259 and the expected number of screened households was 10.5/.259 = 40.5. In an area where 50 per cent of households were black, I = .630, and the expected number of screened households, 16.7. In the first case, the expected cost per interview was raised to (60.49 - .03K) dollars, and in the second case to (32.41 - .03K) dollars where K equals the number of interviews.

Since the same number of white interviews could have been obtained without racial screening by placing all listing areas in Stratum 1, the costs of such screening must be added to the costs of obtaining black interviews. Therefore, the expected cost per black interview would be higher than the overall cost per interview, and the increase would be considerable in areas with few blacks. The estimated costs for areas of different racial composition are presented in Table 4. There we see that the cutoff points between the two strata were about as predicted in Table 1. By subsampling listing areas using the enumerators' estimates of racial composition, the proportion of black eligible respondents in those listing areas of a psu where household screening on the basis of race was carried out never fell below 10 per cent. The actual cost per interview was over $100 in only one psu, and the estimated cost per black interview exceeded $100 in only three psus.

5. Results: Design Effects

The sampling frame used for this study had one unfortunate aspect. The primary sampling units were constant in size, each one being defined to include 10,000 housing units as of the 1960 Census. The geographic area covered by such psus varied greatly. Among the selected psus, the range extended from a psu covering about one square mile on the south side of Chicago to a psu covering over 20,000 square miles in ten counties in eastern Montana. The sample psus were typically small and homogeneous in comparison to psus selected in other frames such as that of the Current Population Survey or the Survey Research Center at the University of Michigan. This difficulty was anticipated before the study began, but because of time constraints, it was not possible to construct a new sampling frame with more heterogenous psus such as the two mentioned above or that now used at ISR.

These characteristics of the psus had two important effects. One was that they were homogeneous, and the values of the intraclass correlation coefficients for study variables were greater than they would have been for other studies. The other was that the black population was concentrated in only a few of the sample psus, so that over half the black interviews were obtained in just 10 of the 126 sample psus. Therefore the design effects, which measure the increase in variance over what would have been ob-

tained from a simple random sample of the same size and is the product of the intraclass correlation and the average cluster size, were increased over what would have been expected even for such homogeneous psus.

Design effects were computed for 12 variables. These are demographic and economic variables for which the design effects are usually larger than for other variables, particularly attitudes. These were computed for four groups, blacks in Stratum 2 subdivided into three groups depending on racial composition, and hence the cost per interview, of the psu, and whites. There were not enough blacks in Stratum 1 to merit separate computations. The three black groups included approximately the same numbers of interviews. These are shown in Table 5.

The intraclass correlations for the blacks were lower than they were for the whites, indicating greater heterogeneity within black areas than white areas. However, the clustering of black interviews was so much greater than the clustering of white interviews that the design effects for blacks were much greater in the two groups of interviews obtained in particularly black areas and nearly as great in the third group, where the cost per interview was much higher.

The lower costs per interview obtained in the areas of greatest black concentration appear to have been obtained at the price of greater design effects. When the costs per interview were multiplied by the design effects, giving the costs per equivalent simple random sampling interview, these latter costs were comparable among the three black groups, ranging from about $100 to $120 per equivalent interview.

The lessons to be learned from this exercise can be summarized as follows:

a. The additional costs of screening, given these costs of the various components of interviewing, become great when the subpopulation comprises about 10 per cent of the total, and rise sharply below that level.

b. When sampling blacks, gains can be made by identifying areas of greater concentration, and applying optimal allocation.

c. However, it is important that the black interviews not be clustered in a few small, homogeneous psus as the design effects will be unduly large.

d. This difficulty can be overcome by having more diverse psus with smaller numbers of black interviews in each. We saw from equation 5 in Table 3 that little reduction in cost per interview is obtained from having many interviews in one psu. However, if areas of greatest black concentration can be identified within psus, these can be sampled at a higher rate, minimizing the amount of screening, spreading the black sample over a greater number of areas, and reducing the number of black interviews in any one psu.

Table 1. Expected Interviewing Costs in Listing Areas Varying in Racial Composition

Percent Black in Listing Area	Estimated Number of Interviews[1] White	Black	Total	Expected Total Hours Spent in Listing Area[2]	Expected Total Hours Required for all White Interviews in Listing Area[3]	Expected Total Hours Required for all Black Interviews in Listing Area	Expected Hours Per Black Interview	Increase in Time Per Black Interview Because Screening Necessary (Square Root)
0	2.78	0	2.78	33.56	12.82	20.74	---	---
1	2.75	.11	2.86	33.72	12.68	21.04	191.27	41.49 (6.44)
2	2.72	.21	2.93	33.86	12.54	21.32	101.52	22.02 (4.69)
2.5	2.71	.27	2.98	33.96	12.49	21.47	79.52	17.25 (4.15)
3	2.69	.32	3.01	34.02	12.40	21.62	67.56	14.66 (3.83)
5	2.64	.54	3.18	34.34	12.17	22.17	41.06	8.91 (2.98)
10	2.50	1.07	3.57	35.14	11.53	23.61	22.07	4.79 (2.19)
25	2.08	2.68	4.76	37.52	9.59	27.93	10.42	2.26 (1.50)
50	1.39	5.36	6.75	41.48	6.41	35.07	6.54	1.42 (1.19)
75	.69	8.04	8.73	45.54	3.18	42.36	5.27	1.14 (1.07)
90	.28	9.64	9.92	47.84	1.28	46.56	4.83	1.05 (1.02)
100	0	10.71	10.71	49.42	--	49.42	4.61	1.00 (1.00)

[1] Assuming 100 households per listing area, 1 household in 7 to include an eligible respondent, and a 75 percent completion rate and blacks selected at 3.86 times the rate of whites.

[2] 3 hours required to list households, 15 minutes to process each household before interviewing, 2 hours to carry out and code each interview.

[3] This figure equals 2 hours times the number of expected white interviews plus 28 hours times (% white/3.86), where 28 hours is estimated time needed to list and contact 100 households.

Table 2. Relative Effects of Four Variables on Accuracy of Estimates of Racial Composition in SMSAs

Equation	Coefficient of Determination	Cases Included
$\hat{Y} = 34.686 - .299X_1 + .282X_2 + 1.357X_3 - .571X_4$	$R^2 = .195$ $r^2_{4y} = .189$	All Estimates n = 253
$\hat{Y} = 35.194 - .109X_1 - .495X_2 + 3.018X_3 - .604X_4$	$R^2 = .257$ $r^2_{4y} = .249$	Estimates Where Error Less than 80% n = 251
$\hat{Y} = 33.665 - .208X_1 - 1.948X_2 + 2.561X_3 - .568X_4$	$R^2 = .386$ $r^2_{4y} = .373$	Estimates Where Error Less Than 50% n = 243

Y = absolute error of estimate,

X_1 = 0 if area located in South, 1 otherwise,

X_2 = 0 if area located in suburb, 1 in central city,

X_3 = 0 if enumerator's race was black, 1 otherwise,

X_4 = absolute difference of actual percentage black from 50 per cent.

Table 3. Regression Equations Estimating Variations in Costs Per Interview Over Primary Sampling Units

1. $\hat{Y}_1 = 2.72 + .72X_1 + .75X_2 + .01X_3$ $R^2 = .485$

2. $\hat{Y}_2 = 3.64 + .21X_1 + .30X_2 - .02X_3$ $R^2 = .145$

3. $\hat{Y}_3 = 6.37 + .93X_1 + 1.03X_2 - .02X_3$ $R^2 = .416$

4. $\hat{Y}_4 = 1.67 + .24X_1 + .33X_2 - .02X_3$ $R^2 = .253$

5. $\hat{Y}_5 = 11.79 + 1.18X_1 + 1.39X_2 - .03X_3$ $R^2 = .443$

X_1 = number of households screened/number of interviews collected,

X_2 = number of nonresponses/number of interviews collected,

X_3 = number of interviews collected.

Y_1 = total salaries paid to interviewers for interviewing/number of interviews collected,

Y_2 = total expenses paid to interviewers for interviewing/number of interviews collected,

Y_3 = total salaries and expenses paid to interviewers for interviewing/number of interviews collected,

Y_4 = total salaries and expenses paid to interviewers for listing/number of interviews collected,

Y_5 = total costs paid for interviewing, listing, and coding/number of interviews collected.

Table 4: Estimated Costs Per Interview and Per Black Interview in Areas Varying in Racial Composition

Percent of all Households which are Black	Percent of Potentially Eligible Respondents not Screened on Basis of Race	Estimated Number of Screened Households per Interview	Cost per Interview[1]	Percent of all Interviews which are with Black Respondents	Cost per Black Interview[2]	Square Root, Cost per Black Interview Divided by Expected Cost per Interview with no Racial Screening
0	25.91	40.5	59.35	0	---	---
1	26.65	39.4	58.05	3.75	933.28	6.24
2	27.39	38.3	56.75	7.30	473.26	4.45
3	28.13	37.3	55.57	10.66	320.54	3.66
4	28.87	36.4	54.51	13.86	244.09	3.19
5	29.61	35.5	53.45	16.89	198.62	2.88
10	33.32	31.5	48.73	30.01	106.52	2.11
20	40.73	25.8	42.00	49.10	60.71	1.59
30	48.13	21.8	37.28	62.33	45.34	1.38
40	55.54	18.9	33.86	72.02	37.71	1.25
50	62.95	16.7	31.27	79.43	33.17	1.18
60	70.36	14.9	29.14	85.28	30.04	1.12
70	77.77	13.5	27.49	90.01	27.88	1.08
80	85.18	12.3	26.07	93.92	26.23	1.05
90	92.59	11.3	24.89	97.20	24.94	1.02
100	100.00	10.5	23.95	100.00	23.95	1.00

[1] Estimated using equation 5, Table 3, setting the number of interviews equal to 38, the mean obtained over the 115 areas.

[2] This assumes that the expected cost per white interview in Stratum 2 equals the expected cost per interview in Stratum 1 equals 25.09-(.03)(38) = $23.95.

Table 5: Design Effects and Relative Costs for Black and White Interviews

Group[1]	Mean, Intraclass Correlation	Average Cluster Size	Mean Design Effect[2]	Cost per Interview	Cost Per Equivalent Simple Random Sampling Interview[3]
Blacks, Set 1	.044	76.2	4.29	23.36	100.21
Blacks, Set 2	.059	52.6	4.02	26.61	106.97
Blacks, Set 3	.043	27.2	2.13	55.90	119.07
Whites	.070	23.7	2.59	23.95	62.05

[1] Blacks, Set 1 includes all black interviews in psus where blacks comprised 90 to 100 percent of population.

Blacks, Set 2 includes all black interviews in psus where blacks comprised 40 to 90 percent of population.

Blacks, Set 3 includes all remaining black interviews in Stratum 2.

Whites includes all white interviews

[2] The design effect was equal to $1 + \text{roh}\,(B-1)$, where roh = intraclass correlation, B = average cluster size

12 variables were used. They are:

- % 8th grade education or less
- % never married
- % yearly household income above $15,000
- % never had intercourse
- % does not pay rent
- % living in owner-occupied house valued under $10,000
- % lived at present address less than 5 years
- % respondents unemployed
- % household heads unemployed
- % never been pregnant
- % intend to have 5 or more children
- % living with parents

[3] This is the product of the cost per interview and the design effect.

BIBLIOGRAPHY

Cochran, William G. Sampling Techniques, Wiley, New York, 1963.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow. Sample Survey Methods and Theory, Vol. I, Wiley, New York, 1953.

Kish, Leslie. Survey Sampling, Wiley, New York, 1964.

Zelnik, Melvin, and John F. Kantner. "Sexuality, Contraception, and Pregnancy in the United States," in U.S. Commission on Population Growth and the American Future, Vol. I, Charles F. Westoff and Robert Parke, Jr., eds., Washington, D.C., Government Printing Office, 1972.

ACKNOWLEDGEMENT

I would like to acknowledge the very capable assistance of Christine Amoroso, who carried out many of the computations presented in this paper.

THE MOMENTS AND DISTRIBUTION OF THE RATIO OF TWO GAMMA VARIABLES

John A. Flueck and Burt S. Holland, Temple University

1. Introduction

We present here work in progress dealing with the ratio of two correlated gamma random variables. Rietz [13] and the authors [4] have presented discussion of some past work in this area.

David and Fix [3] noted that if X, Y, and Z are independent gamma variates with shape parameters a, b, c respectively and common scale parameter λ, e.g.,

$$f_X(x) = (\lambda x)^{a-1}\lambda e^{-\lambda x}/\Gamma(a),\ x > 0,\ a > 0,\ \lambda > 0,$$

then the random variables $U = X + Y$ and $W = X + Z$ are bivariate gamma distributed with density

$$f_{U,W}(u,w) = \frac{e^{-u-w}}{\Gamma(a)\Gamma(b)\Gamma(c)}\int_0^{\min\{u,w\}} t^{a-1}(u-t)^{b-1}(w-t)^{c-1}\cdot e^{t}dt$$

In this paper we shall study the distribution and moments of

$$r' = \frac{X + Y}{X + Z}$$

as an estimator of

$$\frac{E(X + Y)}{E(X + Z)} = \frac{a + b}{a + c}.$$

Since the parameter λ does not appear in the distribution of r' we may henceforth assume without loss of generality that $\lambda = 1$.

The David-Fix formulation of the bivariate gamma distribution constrains the correlation coefficient,

$$\rho = a/[(a+b)(a+c)]^{1/2},$$

to be nonnegative, and thus is somewhat restricted. We are also examining other bivariate gamma distributions (cf. Johnson and Kotz [7]), but since most such distributions are defined in terms of the bivariate characteristic function, the properties of the ratio estimator are more difficult to pursue.

The ratio r' is easily generalized to the ratio of sums of gamma variates,

$$r^* = \sum_{i=1}^{m}(X_i + Y_i)/\sum_{j=1}^{n}(X_j + Z_j), \qquad (1)$$

or the ratio of means, $r = nr^*/m$. We assume mutual independence of all r.v.'s in (1). If we let r'(a, b, c) denote the probability distribution of r', it follows that the distribution of r is related to that of r' according to

$$r \sim \begin{cases} (n/m)\cdot r'(na,(m-n)a+mb,nc), & m>n \\ (n/m)\cdot r'(ma,mb,(n-m)a+nc), & m\le n \end{cases} \qquad (2)$$

2. Moments of Ratios

The mean and variance of r' are readily computed using well-known properties of the gamma and inverted gamma distribution:

$$E(r') = E[X/(X+Z)] + E(Y)\cdot E(X+Z)^{-1}$$

$$= a(a+c)^{-1} + b(a+c-1)^{-1}, \quad a+c>1.$$

If $a + c \le 1$, E(r') is not finite.

$$E(r')^2 = E[X/(X+Z)]^2 + 2E[X/(X+Z)]\cdot E(X+Z)^{-1} \cdot E(Y) + E(Y^2)E(X+Z)^{-2},$$

hence

$$V(r') = \frac{a(a+1)}{(a+c)(a+c+1)} + 2\frac{a}{(a+c)}\frac{1}{(a+c-1)}b + \frac{b^2+b}{(a+c-1)(a+c-2)} - [E(r')]^2,\ a+c>2.$$

For the variance to be finite it is required that a + c exceed 2.

Using (2), the moments of r are easily obtained from those of r':

$$E(r) = \begin{cases} \frac{a}{a+b} + \frac{nb}{n(a+c)-1}, & n(a+c)>1 \text{ and } m \le n \\ \frac{n}{m}\left[\frac{a}{a+c} + \frac{(m-n)a+nb}{n(a+c)-1}\right], & n(a+c)>1 \text{ and } m > n \end{cases}$$

$$V(r) = \begin{cases} (n/m)^2\left\{\frac{ma(ma+1)}{n(a+c)[n(a+c)+1]} + \frac{2m^2ab}{n(a+c)[n(a+c)-1]}\cdot \right. \\ \quad + \frac{m^2b^2+mb}{[n(a+c)-1][n(a+c)-2]} - \left[\frac{ma}{n(a+c)}\right. \\ \quad \left.\left. + \frac{mb}{n(a+c)-1}\right]^2\right\}, & n(a+c) > 2,\ m\le n. \\ (n/m)^2\left\{\frac{na(na+1)}{n(a+c)[n(a+c)+1]} + \frac{2a}{(a+c)}\frac{[(n-m)a+mb]}{n(a+c)-1}\right. \\ \quad + \frac{[(n-m)a+mb]^2+[(n-m)a+mb]}{[n(a+c)-1][n(a+c)-2]} - \\ \quad \left.\left[\frac{a}{(a+c)} + \frac{(m-n)a+nb}{n(a+c)-1}\right]^2\right\} & n(a+c)>2,\ m>n \end{cases}$$

It is interesting to note that the expected ratio does not depend upon the numerator sample size if the numerator sample size is not greater than the denominator sample size.

A particular case that has been examined in detail is that of identically distributed numerator and denominator, and equal sample sizes: $a + b = a + c = A$, $\rho = a/A$, $m = n$. Then

$$E(r) = (nA-\rho)/(nA-1), \quad nA>1 \tag{3}$$

and bias of r as an estimator of $(a+b)/(a+c) = 1$ is

$$\text{Bias } (r;1) = (1-\rho)/(nA-1), \quad nA>1 \tag{4}$$

Also,

$$V(r) = \left[\frac{\rho(nA\rho+1)}{nA+1} + \frac{2nA\rho(1-\rho)}{nA-1} + \frac{nA(1-\rho)+n^2A^2(1-\rho)^2}{(nA-1)(nA-2)} - \frac{(nA-\rho)^2}{(nA-1)^2}\right],$$

$$nA>2. \tag{5}$$

Many textbooks dealing with sampling theory, (see, e.g., Cochran [2]), contain the following Taylor Series approximations for the mean and variance of $r = \bar{y}/\bar{x}$:

$$E(r) \doteq R+R\frac{(1-f)}{n^2}\left[\frac{S(X)^2}{\bar{X}^2} - \rho\,\frac{S(X)}{\bar{X}}\,\frac{S(Y)}{\bar{Y}}\right] \tag{6}$$

$$\text{Bias}(r;R) \doteq R\frac{(1-f)}{n^2}\left[\frac{S(X)^2}{\bar{X}^2} - \rho\frac{S(X)}{\bar{X}}\,\frac{S(Y)}{\bar{Y}}\right] \tag{7}$$

$$V(r) \doteq R^2\,\frac{(1-f)}{n^2}\left[\frac{S(X)^2}{\bar{X}^2} + \frac{S(Y)^2}{\bar{Y}^2} - 2\rho\frac{S(X)}{\bar{X}}\,\frac{S(Y)}{\bar{Y}}\right] \tag{8}$$

Here $R = \bar{Y}/\bar{X}$, $\bar{Y}$, and $\bar{X}$ are the population means of Y and X, S(X), and S(Y) are the population standard deviations of X and Y, ρ is the population product moment correlation coefficient between X and Y, and $f = n/N$, the ratio of sample to population size. As is well-known, the expectation bias can be sizable.

If we apply (6-8) to the David-Fix moments (3), (5), assuming $f = 0$, we obtain

$$E(r) \doteq (n^2A - \rho+1)/n^2A \tag{9}$$

$$\text{Bias } (r;1) \doteq (1-\rho)/n^2A \tag{10}$$

$$V(r) \doteq 2(1-\rho)/n^2A \tag{11}$$

Comparing the Taylor Series expansion results (6-8) with the exact results (9-11) we see that the conventionally-used Taylor Series Approximations underestimate the expectation bias and variance of r by roughly a multiplicative factor of n.

3. Some Empirical Results

To illustrate the impact of this underestimation, we present examples for two carefully documented weather modification experiments, [5], [14]. Barger and Thom [1] and others have noted that rainfall data are often well-fitted by the gamma distribution. Meteorologists often report the efficacy of cloud seeding experiments via the use of a double ratio, for example,

$$DR = \frac{\text{area one seeded mean}}{\text{area two unseeded mean}} \Big/ \frac{\text{area two seeded mean}}{\text{area one unseeded mean}},$$

the product of two independent ratios of means, or its square root (see [9]). Table 1 contains the reported square root of double ratio (RDR); the RDR with each of the ratios corrected for expectation bias using the exact correction (4); and the large sample approximation to the variance of the product of two independent ratios (first reported by J.N.K. Rao[12]) using both the Taylor series approximation for V(r), (8), and the exact value for V(r), (5). It is seen that the RDR corrections are at times appreciable and that the approximate variance of the double ratios employing the exact V(r) is much larger than if the approximate V(r) had been employed.

In reference to their experience based on a considerable number of sampling experiments, and using the Taylor Series approximation for bias and standard error of r, Kish, Namboodiri, and Pillai [8] state, "the ratio of bias of r to standard error of r averaged around .01 and seldom appears greater than .04". In Table 2, we have computed the exact bias to standard error ratio using the exact moments (3), (5) for selected values of ρ, n and A. The selected values are not atypical of rainfall data and may be applicable elsewhere. At least for those parameter values tabulated, the exact bias to standard error ratio is not consistent with the Kish et al statement.

4. Densities

We have recently begun work on the exact probability density of r' (from which the exact probability density of r follows). If we let $V = X/(X+Z)$ and $W = X+Z$, the joint density of r', V and W is easily shown to be

$$g(r',v,w) = \frac{(vw)^{a-1}e^{-vw}}{\Gamma(a)} \cdot \frac{[w(r'-v)]^{b-1}e^{-w(r'-v)}}{\Gamma(b)} \cdot \frac{[w(1-v)]^{c-1}e^{-w(1-v)}}{\Gamma(c)} \cdot w^2,$$

$$0 < w$$
$$0 < r'$$
$$0 < v < \min\{1,r'\}$$

Then the joint density of r' and v is

$$g(r',v) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} \cdot \frac{v^{a-1}(r'-v)^{b-1}(1-v)^{c-1}}{(1+r'-v)^{a+b+c}}, \quad 0 < v < \min\{1,r'\}, \; 0 < r'$$

and the p.d.f. of r' is

$$g(r') = \int_0^{\min\{1,r'\}} g(r',v)dv, \qquad r' > 0 \qquad (12)$$

In the case where X, Y, and Z are exponential random variables (a=b=c=1), this becomes,

$$g(r') = \begin{cases} 1 - 1/(r'+1)^2, & 0 < r' \le 1 \\ 1/(r')^2 - 1/(r'+1)^2, & r' > 1. \end{cases} \qquad (13)$$

A closed form expression for (12) appears to be readily obtainable only for small integral values of a, b, c. We have also tried to obtain g(r') via an inversion formula in Gurland [6] that uses the joint characteristic function of X, Y and Z, but the integral in the inversion formula has thus far proved elusive.

Instead, a computer program has been developed for the numerical integration needed to generate and plot g(r') and the c.d.f. of r'. Depending on the parameter values under study, a variety of shapes are possible--some sketches of p.d.f.'s appear in Figure 1. These results exhibit both unimodal and bimodal densities with the possibility of an asymptote at one.

The main determinant of the shape of g(r') is the value of b + c. When b + c > 1, the distribution is unimodal and finite for all r' > 0. If b + c $\le$ 1,

$$\lim_{r' \to 1} g(r') = \infty.$$

Note that in the distribution of r with a + b = a + c = A, ρ = a/A, and m = n, b + c corresponds to the quantity 2nA(1-ρ).

Figure 1(a) displays the p.d.f. in (13), i.e., a = b = c = 1. g(r') has a discontinuous derivative at r' = 1. In Figure 1(b), g(r') has two inflection points below the modal value 1.30. The expected value of both numerator and denominator is larger for the example in Figure 1(c) than in any of the other cases. In the distribution of r this would correspond to a large sample size if A and ρ are small. C.R. Rao [11] has shown that under rather general conditions, the distribution of the ratio of two means is asymptotically normal. This explains the relative lack of shewness in 1(c). In Figure 1(d), the graph of g(r') is monotonically increasing to the left of 1. However, in Figure 1(e), the distribution has a mode between 0 and 1. This might be viewed by some as a bimodal distribution. These results are somewhat similar to the findings of Marsaglia [10] that the distribution of the ratio of two correlated normal random variables (with his assumed bivariate structure) may be bimodal as well as unimodal.

Needless to say, the distribution of r' and probably r is not necessarily well-behaved, and the automatic reliance on the Central Limit Theorem to assure normality for moderate sized samples is highly questionable. Further work is under way.

5. Acknowledgement

The authors wish to acknowledge the assistance of Ru-Ying Lee, Yong Soo Kim, and the Temple University Computer Activity in generating the computer plots.

References

[1] Barger, G.L. and Thom, H.C.S. (1949). "Evaluation of Drought Hazard," Agron. Journ. 41, 519-526.

[2] Cochran, W.G. (1963). Sampling Techniques, Second Edition. New York: John Wiley and Sons, Inc.

[3] David, F.N. and Fix, E. (1961). "Rank Correlation and Regression in a Nonnormal Surface," Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, 177-197.

[4] Flueck, J. and Holland, B. (1973). "Ratio Estimation Problems in Meteorological Research." Proceedings, Third American Meteorological Society Conference on Probability and Statistics in Atmospheric Science. Boulder, Colo., June 1973.

[5] Godson, W.L., Crozier, C.L. and Holland, J.D. (1966). "An Evaluation of Silver-Iodide Cloud Seeding by Aircraft in West Quebec, Canada, 1961-63," Journ. Appl. Meteorology, 5, 500-512.

[6] Gurland, J. (1948). "Inversion Formulae for the Distribution of Ratios," Annals of Math. Stat., 19, 1948.

[7] Johnson, N.L. and Kotz, S. (1972). Distributions in Statistics: Continuous Multivariate Distributions. New York: John Wiley and Sons, Inc.

[8] Kish, L., Namboodiri, N.K., and Pillai, R.K. (1962). "The Ratio Bias in Surveys," J. Amer. Statist. Assoc. 57, 863-876.

[9] LeCam, L. and Neyman, J. (eds.) (1967). Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. V, Weather Modification. Berkeley: University of California Press.

[10] Marsaglia, G. (1965). "Ratios of Normal Variables and Ratios of Sums of Uniform Variables," J. Amer. Statist. Assoc., 60, 193-204.

[11] Rao, C.R. (1952). Advanced Statistical Methods in Biometric Research. New York: John Wiley and Sons, Inc.

[12] Rao, J.N.K. (1957). "Double Ratio Estimate in Forest Surveys," Journ. Indian Soc. Agr. Stat., 9, 191-204.

[13] Rietz, H.L. (1936). "On the Frequency Distribution of Certain Ratios," Ann. Math. Statist. 7, 145-153.

[14] Smith, E.J. (1967). "Cloud Seeding Experiments in Australia," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. V, 161-76. Univ. of California Press.

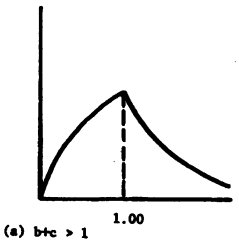
Table 1. Environmental Results from Cloud Seeding Experiments

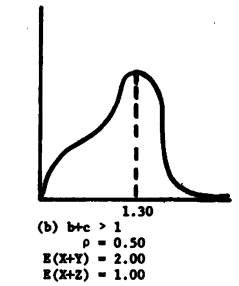
Project		n	RDR	CRDR	AVDR1	AVDR2
West Quebec	1960-63	23	.9310	.9154	.0022	.0566
	1960	6	.9275	.8560		
	1961	5	.6801	.6182		
	1962	5	1.793	1.695		
	1963	7	.8618	.8127		
South Australia	1957-59	22	.9525	.9473	.0008	.0174
	1957	8	.9711	.9563		
	1958	9	.9234	.9168		
	1959	5	.8836	.8482		

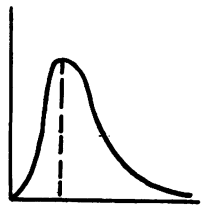
Key:
RDR = square root of product of two independent ratios
CRDR = RDR with each of ratios corrected for expectation bias using $(1 - \rho)/(nA - 1)$
AVDR1 = large sample approximation to variance of product of two independent ratios, [12], inserting $V(r) = 2(1 - \rho)/n^2A$ in the approximation
AVDR2 = same as AVDR1 except using exact value for V(r)

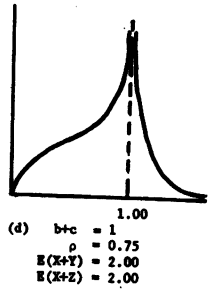
Table 2. Relative Bias of r for Selected Values of ρ, n, A

ρ	n	A	bias/σ
.00	10	.25	0.224
.00	10	2.0	0.152
.00	50	.25	0.187
.00	50	2.0	0.0702
.50	10	.25	0.188
.50	10	2.0	0.110
.50	50	.25	0.137
.50	50	2.0	0.0505
.95	10	.25	0.0756
.95	10	2.0	0.0358
.95	50	.25	0.0454
.95	50	2.0	0.0158









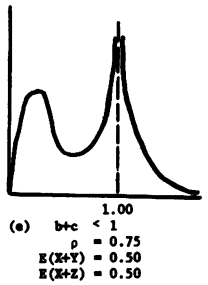


Figure 1. Selected Sketches of Computer Plots of g(r')

EMPIRICAL RESULTS IN FINITE POPULATION SAMPLING

Richard L. Greenstreet. Virginia Polytechnic Institute and State University

A. Statement of Problem

i. Given: A finite population of units $U_1,\ldots,U_N$ with $y_1,\ldots,y_N$ unknown before sampling and $x_1,\ldots,x_N$, constants, known before sampling. We wish to estimate

$$T = \sum_{j=1}^{N} y_j .$$

ii. How to Estimate T

a. In the traditional theory (see, for example, Horwitz and Thompson (1952)) one takes a simple random sample of size n and uses, perhaps,

$$\frac{N}{n}\sum_s y \quad \text{or} \quad \frac{\sum_s y}{\sum_s x}\sum_{i=1}^{N} x_i .$$

b. Prediction Theory (Least Squares Approach) $y_1,\ldots,y_N$ are realized values of random variables $Y_1,\ldots,Y_N$. $E(Y_i)=f(x_i)$ and $\mathrm{Var}(Y_i)=\sigma^2 v(x_i)$. For example, if $E(Y_i)=\beta x_i$, $\mathrm{Var}(Y_i)=\sigma^2 x_i$, then the linear estimator, $\hat{T}$ for $T=\sum_{j=1}^{N} Y_j$ which minimizes $E(T-\hat{T})^2$ and satisfies $E(T-\hat{T})=0$ is $\hat{T}=\frac{\sum_s y}{\sum_s x}\sum_{i=1}^{N} x_i$, i.e., the ratio estimator.

iii. Bias Within the Prediction Context

Within the prediction approach to estimation in finite populations an estimator $\hat{T}$ for T is said to be unbiased for T if $E(\hat{T}-T)=0$. Now, under a model such that $E(Y_i)=\beta_0+\beta_1 x_i$, we have

$$E\left(T-\frac{\sum_s y}{\sum_s x}\sum_{i=1}^{N} x_i\right)=\beta_0\left(\frac{N\sum_s x_i - n\sum_{i=1}^{N} x_i}{\sum_s x_i}\right)$$

which is, in general, nonzero unless

$$\frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_s x_i}{n},$$

that is, unless the units are chosen in such a way that the average on the auxiliary variable x within the sample is equal to the average on x within the entire finite population. Such a sample is said to be balanced on the first moment of x.

B. Objective of Paper

It is the aim of this paper to demonstrate the importance of balanced samples. This will be accomplished by showing that the error in estimating a known total for a specific population is much smaller when the samples are balanced than when samples departing significantly from balance are used. It will also be noted that measures of error for the estimate of the total will have more validity when balanced samples are used. This result holds for a wide variety of estimators.

C. The Population and Sampling Procedure

The population studied consisted of 150 Negro males, ages 10, 11, and 12. On each individual the three variables, weight, height, and chest circumference were measured. Samples of sizes 5, 10, and 30 were selected.

For each sample size, samples composed of the largest and smallest units on the variable height as well as samples approximately balanced on height were chosen. The balanced samples were obtained by ranking the units on height and selecting the proper number of units equally spaced on the ranks of height. For example, a sample approximately balanced on height of size 30 was obtained by selecting the 5th largest, 10th largest, 15th largest, etc., down to the smallest unit on the variable height. A variety of estimators were compared with regard to their ability to estimate the total weight of the 150 Negro males. For each estimator studied, the associated estimate of variance was also tabulated.

The Estimators

The estimators will now be listed. Some of the estimators use two auxiliary variables, x_0 and x_1. For this study the variable weight played the role of y, height the role of x_1, and chest circumference the role of x_0.

Olkin's Estimator

Olkin's estimator with two auxiliary variables x_0 and x_1 is given by

$$\hat{T}_{OL} = \omega_1 \frac{\sum_s y_i}{\sum_s x_{j1}} \sum_{j=1}^{N} x_{j1} + \omega_2 \frac{\sum_s y_j}{\sum_s x_{j0}} \sum_{j=1}^{N} x_{j0} \qquad (1)$$

where $\omega_1+\omega_2=1$. The weights ω_1,ω_2 depend upon the sample through a 2x2 matrix A, where

$$a_{11}=\sum_s\left(y_j-\frac{\sum_s y_i}{\sum_s x_{j1}}x_{j1}\right)^2,\quad a_{22}=\sum_s\left(y_j-\frac{\sum_s y_i}{\sum_s x_{j0}}x_{j0}\right)^2 \text{ and}$$

$$a_{12}=a_{21}=\sum_s\left(y_j-\frac{\sum_s y_i}{\sum_s x_{j1}}x_{j1}\right)\left(y_j-\frac{\sum_s y_j}{\sum_s x_{j0}}x_{j0}\right).$$

Then $\underset{\sim}{\omega}=(\omega_1,\omega_2)=\frac{e'A^{-1}}{e'A^{-1}e}$ where $e'=(1,1)$. A measure of uncertainty for (1) is given by

$$\hat{\sigma}_{OL}=\left[\frac{N^2}{ne'A^{-1}e}\right]^{1/2}. \qquad (2)$$

Least Squares Prediction Estimators

Now, under the model $Y_j=\beta_2 x_{j1}+\varepsilon_{j2}[v_2(x_{j1})]^{1/2}$ where β_2 is unknown but constant, $E(\varepsilon_{j2})=0$, and and $\mathrm{VAR}(\varepsilon_{j2})=\sigma_2^2$; and an analogous model $Y_j=\beta_1 x_{j0}+\varepsilon_{j1}[v_1(x_{j0})]^{1/2}$ where again β_1 is unknown but constant, $E(\varepsilon_{j1})=0$, and $\mathrm{VAR}(\varepsilon_{j1})=\sigma_1^2$, the least squares theory of estimation in finite populations gives as the optimal estimator when $v_2(x_{j1})=x_{j1}$ and $v_1(x_{j0})=x_{j0}$ just the ratio estimators based on x_1 and x_0 respectively. Expressions (3) - (8) list the appropriate estimators under these two models along with measures of uncertainty provided by least squares theory and traditional finite sampling theory. Let

$$\bar{x}_i(\bar{s})=\frac{1}{N-n}\sum_{\bar{s}} x_{ji},\quad \bar{x}_i(s)=\frac{1}{n}\sum_s x_{ji} \text{ and } \bar{x}_i(P)=\frac{1}{N}\sum_{j=1}^{N} x_{ji},\ i=0,1.$$

Then

$$R(x_1)=\frac{\sum_s y_j}{\sum_s x_{j1}}\sum_{j=1}^{N} x_{j1}, \qquad (3)$$

$$\hat{\sigma}_{R(x_1)} = \left[\frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_s\left(y_j - \frac{\sum_s y_j}{\sum_s x_{j1}} x_{j1}\right)^2/x_{j1}\right. \quad (4)$$
$$\left. \cdot \frac{\bar{x}_1(\tilde{s})}{\bar{x}_1(s)} \bar{x}_1(P)\right]^{1/2} \quad \text{(least squares)},$$

$$\hat{\sigma}_{R(x_1)} = \left[\frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_s\left(y_j - \frac{\sum_s y_j}{\sum_s x_{j1}} x_{j1}\right)^2\right]^{1/2} \quad (5)$$
(traditional),

$$R(x_0) = \frac{\sum_s y_j}{\sum_s x_{j0}} \sum_{j=1}^{N} x_{j0}, \quad (6)$$

$$\hat{\sigma}_{R(x_0)} = \left[\frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_s\left(y_j - \frac{\sum_s y_j}{\sum_s x_{j0}}\right)^2/x_{j0}\right. \quad (7)$$
$$\left. \cdot \frac{\bar{x}_0(\tilde{s})}{\bar{x}_0(s)} \bar{x}_0(P)\right]^{1/2} \quad \text{(least squares)},$$

$$\hat{\sigma}_{R(x_0)} = \left[\frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_s\left(y_j - \frac{\sum_s y_j}{\sum_s x_{j0}}\right)^2\right]^{1/2} \quad (8)$$
(traditional).

When $v_2(x_{j1}) = v_1(x_{j0}) = 1$, least squares gives the following estimators and measures of uncertainty:

$$\hat{T}_1 = \sum_s y_j + \hat{\beta}_2 \sum_{\tilde{s}} x_{j1} \text{ where } \hat{\beta}_2 = \frac{\sum_s y_j x_{j1}}{\sum_s x_{j1}^2} \quad (9)$$

$$\hat{\sigma}_{\hat{T}_1} = \left[\hat{\sigma}_2^2\left[\frac{\left(\sum_{\tilde{s}} x_{j1}\right)^2}{\sum_s x_{j1}^2} + (N-n)\right]\right]^{1/2} \quad (10)$$

where $\hat{\sigma}_2^2 = \frac{1}{n-1}\sum_s (y_j - \hat{\beta}_2 x_{j1})^2$.

$$\hat{T}_2 = \sum_s y_j + \hat{\beta}_1 \sum_{\tilde{s}} x_{j0}, \quad (11)$$

$$\hat{\sigma}_{\hat{T}_2} = \left[\hat{\sigma}_1^2\left[\frac{\left(\sum_{\tilde{s}} x_{j0}\right)^2}{\sum_s x_{j0}^2} + (N-n)\right]\right]^{1/2} \quad (12)$$

where $\hat{\beta}_1 = \frac{\sum_s y_j x_{j0}}{\sum_s x_{j0}^2}$ and $\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_s (y_j - \hat{\beta}_1 x_{j0})^2$.

Another least squares model whose inclusion is quite natural is

$y_j = \beta_1 x_{j0} + \beta_2 x_{j1} + \epsilon_j [v(x_{j0}, x_{j1})]^{1/2}$. The variance function $v(x_{j0}, x_{j1}) = 1$ was specified, not through any reasoning process, in fact, but by default, since the variation of y over the x_0, x_1 plane is difficult to determine. Under this model, we have through least squares theory,

$$\hat{T}_3 = \sum_s y_j + \hat{\beta}_1 \sum_{\tilde{s}} x_{j0} + \hat{\beta}_2 \sum_{\tilde{s}} x_{j1} \text{ where} \quad (13)$$

$$\hat{\beta}_1 = \frac{1}{\sum_s x_{j0}^2 \sum_s x_{j1}^2 - \left(\sum_s x_{j1} x_{j0}\right)^2}$$
$$\cdot \left[\sum_s x_{j1}^2 \sum_s y_j x_{j0} - \sum_s x_{j1} x_{j0} \sum_s y_k x_{j1}\right]$$

$$\hat{\beta}_2 = \frac{1}{\sum_s x_{j0}^2 \sum_s x_{j1}^2 - \left(\sum_s x_{j1} x_{j0}\right)^2}$$
$$\cdot \left[\sum_s x_{j0}^2 \sum_s y_j x_{j1} - \sum_s x_{j1} x_{j0} \sum_s y_j x_{j0}\right]$$

and a measure of uncertainty is

$$\hat{\sigma}_{\hat{T}_3} = \left[\hat{\sigma}^2\left[\frac{\sum_s (x_{j1} \sum_{\tilde{s}} x_{j0} - x_{j0} \sum_{\tilde{s}} x_{j1})^2}{\sum_s x_{j0}^2 \sum_s x_{j1}^2 - \left(\sum_s x_{j1} x_{j0}\right)^2} + (N-n)\right]\right]^{1/2} \quad (14)$$

where $\hat{\sigma}^2 = \frac{1}{n-2}\left[\sum_s (y_j - \hat{\beta}_1 x_{j0} - \hat{\beta}_2 x_{j1})^2\right]$.

Finally, since the expansion estimator plays such an important role in sampling theory, being a favorite traditional estimator and the optimal estimator in the least squares theory when the sampling is balanced, it was included along with a measure of uncertainty specified by the least squares model $Y_j = \mu + \epsilon_j$, under which it is optimal for any sample.

$$\hat{T}_4 = \frac{N}{n}\sum_s y_j \quad (15)$$

$$\hat{\sigma}_{\hat{T}_4} = \left[\frac{N(N-n)}{n(n-1)}\sum_s (y_j - \bar{y}(s))^2\right]^{1/2} \quad (16)$$

Srivastava's Estimators

Srivastava [1971] proposes a class of estimators, three of which are considered in this study. Each estimator is an adjusted expansion estimator and they are given by (17) - (19)

$$\hat{T}_1 = \frac{N}{n}\left(\sum_s y_j\right) \exp\{\sum_{i=0}^{1} \theta_i \log u_i\} \quad (17)$$

$$\hat{T}_2 = \frac{N}{n}\left(\sum_s y_j\right) \exp\{\sum_{i=0}^{1} \theta_i (u_i - 1)\} \quad (18)$$

$$\hat{T}_3 = \frac{N}{n}\left(\sum_s y_j\right)\{\sum_{i=0}^{1} \omega_i \exp\{(\theta_i/\omega_i) \log u_i\}\}. \quad (19)$$

where $\omega_0 + \omega_1 = 1$, $u_i = \frac{\bar{x}_i(s)}{\bar{x}_i(P)}$ $i = 0, 1$ and

$$\underset{\sim}{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = -A^{-1} \underset{\sim}{b}.$$

In this study the values $\omega_0 = \omega_1 = 1/2$ were selected. For two auxiliary variables A has elements

$$a_{11} = \frac{\sum_s (x_{j0} - \bar{x}_0(s))^2 (\bar{y}(s))^2}{(\bar{x}_0(s))^2 \sum_s (y_j - \bar{y}(s))^2},$$

$$a_{22} = \frac{\sum_s (x_{j1} - \bar{x}_1(s))^2 (\bar{y}(s))^2}{(\bar{x}_1(s))^2 \sum_s (y_j - \bar{y}(s))^2} \quad \text{and}$$

$$a_{12} = a_{21} = \frac{\sum_s (x_{j0} - \bar{x}_0(s))(x_{j1} - \bar{x}_1(s))(\bar{y}(s))^2}{\bar{x}_0(s)\,\bar{x}_1(s) \sum_s (y_j - \bar{y}(s))^2}$$

The vector $\underset{\sim}{b}$ has elements

$$b_1 = \frac{\sum_s (y_j - \bar{y}(s))(x_{j0} - \bar{x}_0(s))\,\bar{y}(s)}{\sum_s (y_j - \bar{y}(s))^2\,\bar{x}_0(s)},$$

$$b_2 = \frac{\sum_s (y_j - \bar{y}(s))(x_{j1} - \bar{x}_1(s))\bar{y}(s)}{\sum_s (y_j - \bar{y}(s))^2 \bar{x}_1(s)} .$$

A measure of uncertainty for (17) - (19) is given by

$$\hat{\sigma}_S = [\frac{N(N-n)}{n(n-1)} \sum_s (y_j - \bar{y}(s))^2 (1 - \underset{\sim}{b}' A^{-1} \underset{\sim}{b})]^{1/2}. \quad (20)$$

(See page 405 of Srivastava [1971].)

Singh's Estimators

We consider four estimators discussed by Singh [1967]. The estimators along with their measures of uncertainty are given by (21) - (28).

$$\hat{T}_1 = \frac{N}{n}(\sum_s y_j) \frac{\sum_s x_{j1} \sum_1^N x_{j0}}{\sum_1^N x_{j1} \sum_s x_{j0}} \quad (21)$$

$$\hat{\sigma}_{\hat{T}_1} = [\frac{(N-n)}{Nn} (\sum_{j=1}^N y_j)^2 [C_y^2 + C_{x_0}^2 - 2C_y C_{x_0} \rho_{y,x_0} + C_{x_1}^2 + 2C_y C_{x_1} \rho_{y,x_1} - 2C_{x_0} C_{x_1} \rho_{x_0,x_1}]]^{1/2} \quad (22)$$

$$\hat{T}_2 = \frac{N}{n} (\sum_s y_j) \frac{\sum_s x_{j0} \sum_1^N x_{j1}}{\sum_1^N x_{j0} \sum_s x_{j1}} \quad (23)$$

$$\hat{\sigma}_{\hat{T}_2} = [\frac{(N-n)}{Nn} (\sum_{j=1}^N y_j)^2 [C_y^2 + C_{x_1}^2 - 2C_y C_{x_1} \rho_{y,x_1} + C_{x_0}^2 + 2C_y C_{x_0} \rho_{y,x_0} - 2C_{x_0} C_{x_1} \rho_{x_0,x_1}]]^{1/2} \quad (24)$$

$$\hat{T}_3 = \frac{n}{N} (\sum_s y_j) \frac{\sum_1^N x_{j0} \sum_1^N x_{j1}}{\sum_s x_{j0} \sum_s x_{j1}} \quad (25)$$

$$\hat{\sigma}_{\hat{T}_3} = [\frac{(N-n)}{Nn} (\sum_{j=1}^N y_j)^2 [C_y^2 + C_{x_0}^2 - 2C_y C_{x_0} \rho_{y,x_0} + C_{x_1}^2 - 2C_y C_{x_1} \rho_{y,x_1} + 2C_{x_0} C_{x_1} \rho_{x_0,x_1}]]^{1/2} \quad (26)$$

$$\hat{T}_4 = (\frac{N}{n})^3 \frac{\sum_s y_j \sum_s x_{j0} \sum_s x_{j1}}{\sum_1^N x_{j0} \sum_1^N x_{j1}} \quad (27)$$

$$\hat{\sigma}_{\hat{T}_4} = [\frac{(N-n)}{Nn} (\sum_{j=1}^N y_j)^2 [C_y^2 + C_{x_1}^2 + 2C_y C_{x_1} \rho_{y,x_1} + C_{x_0}^2 + 2C_y C_{x_0} \rho_{y,x_0} + 2C_{x_0} C_{x_1} \rho_{x_0,x_1}]]^{1/2}$$

For each of the measures of uncertainty $(\sum_{j=1}^N y_j)^2$ was estimated by the square of the expansion estimator. A term like C_y^2 was estimated by

$$\hat{C}_y^2 = \frac{n^2}{(n-1)} \frac{\sum_s (y_j - \bar{y}(s))^2}{(\sum_s y_j)^2}$$

and ρ with any pair of subscripts by the appropriate Pearson produce moment correlation coefficient based on observations in the sample.

Discussion

Tables I and II list certain parameters of the population and the samples chosen from the population. Note, for example, how closely the sample mean $\bar{h}(s)$ for samples 9-13 approximates the population mean of 57.17 for the variable height. For extreme samples, for example 1 and 7, the balance is not so good. Table III gives actural error in estimating the known total of 12416 lbs. along with the ratio of absolute error to estimated standard error for Sriviastava's extimators. For example, column 3 of table III gives 21.9 as the ratio of the absolute error, 5428, to the calculated value of expression (20), for sample I. Note the inadequacy of (20), as a measure of true error in this case. The results on standard error for Singh's estimators are similar to those observed in table III. Tables IV -VI compare balanced samples with samples based on the extreme units of the variables height.

Summary of Results

For samples departing from balance, that is, those based on the largest and smallest units, quite poor estimates of the true total and unrealistic measures of error are observed, particularly for small sample sizes.

The estimators of Singh and Srivastava performed poorly for samples based on the extreme units and many of their estimators perform poorly on balanced samples. The error estimates for Singh's and Srivastava's estimators, in general, perform well only for balanced samples.

The results suggest that sampling plans which insure balance are preferred to those under which extreme departures from balance are possible such as simple random sampling.

TABLE I

PARAMETERS OF POPULATION WHERE N IS THE TOTAL NUMBERS OF UNITS AND h, w, AND c, DESIGNATE HEIGHT, WEIGHT AND CHEST CIRCUMFERENCE RESPECTIVELY

N =150

$\sum_1^N w_j$=12416 lbs.

$\sum_1^N h_j$=8575 ins., $\sum_1^N h_j/150$=57.17, $\sum_1^N h_j^2/150$=3276

$\sum_1^N c_j$=3998 ins., $\sum_1^N c_j/150$=26.66, $\sum_1^N c_j^2/150$= 715

TABLE II

SAMPLE NUMBER AND SIZE - SAMPLING CONFIGURATION AND FIRST AND SECOND SAMPLE MOMENTS FOR THE TWO AUXILIARY VARIABLES

Sample Number	n	Sampling Confuguration	$\bar{h}(s)$	$\bar{h}^2(s)$	$\bar{c}(s)$	$\bar{c}^2(s)$
1	5	1,. . . ,5	64.2	4123	28.6	817
2	5	10,40,...130	57.6	3324	26.7	713
3	5	16,46,..,136	57.2	3274	26.5	715
4	5	15,45,..,135	57.2	3274	26.5	707
5	5	14,44,..,134	57.2	3282	26.9	730
6	5	20,50,..,140	56.8	3235	25.5	652
7	5	146, . .,150	51.0	2603	25.4	644
8	10	1,. . ., 10	63.1	3985	28.1	792
9	10	5,20,..,150	57.4	3304	26.3	694
10	10	7,22,..,140	57.3	3287	25.6	661
11	10	8,23,..,143	57.2	3280	26.6	710
12	10	9,24,..,144	57.1	3266	28.5	823
13	10	10,24,..,145	57.0	3258	26.5	704
14	10	141, . .,150	51.9	2698	25.2	636
15	30	1,. . ., 30	61.3	3759	28.1	796
16	30	5,10,..,150	60.0	3253	26.4	701
17	30	1,6,. .,146	57.4	3301	26.8	726
18	30	2,7,. .,147	57.3	3288	25.9	676
19	30	3,8,. .,148	57.2	3276	26.7	713
20	30	4,9,. .,149	57.1	3263	27.4	757
21	30	121, . .,150	53.3	2846	25.6	658

TABLE III

ACTURAL ERROR AND ESTIMATE OF STANDARD ERROR ALONG WITH THE RATIO OF ABSOLUTE ERROR TO ESTIMATED STANDARD ERROR

			Srivastava's Estimators				
	(17)		(18)		(19)		(20)
Sample Number	$\hat{T} - T$	$\lvert\hat{T} - T\rvert/\widehat{S.E.}$	$\hat{T} - T$	$\lvert\hat{T} - T\rvert/\widehat{S.E.}$	$\hat{T} - T$	$\lvert\hat{T} - T\rvert/\widehat{S.E.}$	$\widehat{S.E.}$
1	- 5428	21.9	- 5750	23.2	- 5105	20.6	248
2	- 756	3.8	- 757	3.5	- 765	3.6	217
3	+ 3.3	.0	+ 3.3	.0	+ 4.0	.0	124
4	+ 11.9	.1	+ 11.7	.1	+ 12.3	.1	101
5	+ 440	10.7	+ 439	10.7	+ 441	10.7	41
6	+ 126	.5	+ 106	.4	+ 149	.8	246
7	- 5285	18.7	- 2138	18.2	- 4168	14.8	282
8	- 1942	4.6	- 2145	5.0	- 1597	3.7	426
9	+ 391	2.0	+ 388	1.9	+ 399	2.0	200
10	- 393	1.4	- 400	1.4	- 386	1.3	291
11	+ 198	.5	+ 195	.5	+ 195	.9	381
12	- 275	.5	- 342	.6	- 108	.2	536
13	- 656	2.5	- 657	2.5	- 656	2.5	260
14	- 1702	6.4	- 1740	6.4	- 1697	6.4	264
15	- 939	4.3	- 1042	4.8	- 927	4.3	217
16	- 223	1.4	- 224	1.4	- 222	1.4	157
17	- 80	.5	- 80	.5	- 80	.5	171
18	- 71	.3	- 75	.4	- 64	.3	212
19	+ 185	.8	+ 185	.8	+ 185	.8	221
20	- 256	1.1	- 267	1.1	- 227	1.0	237
21	+ 450	3.3	+ 375	2.8	+ 483	3.6	136

TABLE IV

AVERAGE ABSOLUTE ERROR FOR BALANCED SAMPLES AND ABSOLUTE ERROR FOR EXTREME SAMPLES BASED ON LARGEST AND SMALLEST UNITS FOR EACH ESTIMATOR WITHIN EACH SAMPLE SIZE. COUNT OF ABSOLUTE ERRORS FALLING WITHIN TWO STANDARD ERRORS, UNDER "VARIANCE EVALUATION"

n	Sample Type	Least Squares (9)	Least Squares (11)	Srivastava (17)	Srivastava (18)
5	BAL	438	333	268	263
	EX L	2970	3709	5428	5750
	EX S	1581	2233	5285	5138
10	BAL	781	583	382	396
	EX L	2050	2728	1942	2145
	EX S	1792	2201	1702	1740
30	BAL	357	200	163	166
	EX L	1792	1995	932	1042
	EX S	1098	1419	450	375
Variance Evaluation		14	15	12	12

TABLE V

AVERAGE ABSOLUTE ERROR FOR BALANCED SAMPLES AND ABSOLUTE ERROR FOR EXTREME SAMPLES BASED ON LARGEST AND SMALLEST UNITS FOR EACH ESTIMATOR WITHIN EACH SAMPLE SIZE. COUNT OF ABSOLUTE ERRORS FALLING WITHIN TWO STANDARD ERRORS, UNDER "VARIANCE EVALUATION"

n	Sample Type	Ratio Aux (h)	Ratio Aux (c)	Ratio Aux (c^2)	Ratio Aux (ch)
5	BAL	476	334	281	335
	EX L	2968	3697	2691	1956
	EX S	1582	2250	1685	1012
10	BAL	756	515	343	495
	EX L	2034	2711	1987	1306
	EX S	1789	2203	1561	1157
30	BAL	350	181	178	173
	EX L	1719	1935	1189	989
	EX S	1106	1436	953	632
Variance Evaluation		15 (5) 14 (4)	15 (5) 15 (4)	15	15

n	Sample Type	Ratio Aux c^2h	Olkin's (1)	Least Squares (13)
5	BAL	313	485	506
	EX L	1082	3789	3796
	EX S	362	3094	3096
10	BAL	317	521	517
	EX L	669	2592	2614
	EX S	430	2222	2221
30	BAL	192	225	212
	EX L	313	2090	2112
	EX S	97	1541	1530
Variance Evaluation		19	14	18

TABLE VI

AVERAGE ABSOLUTE ERROR FOR BALANCED SAMPLES AND ABSOLUTE ERROR FOR EXTREME SAMPLES BASED ON LARGEST AND SMALLEST UNITS FOR EACH ESTIMATOR WITHIN EACH SAMPLE SIZE. COUNT OF ABSOLUTE ERRORS FALLING WITHIN TWO STANDARD ERRORS, UNDER "VARIANCE EVALUATION"

n	Sample Type	Srivastava (19)	Expansion (15)	Singh (21)
	BAL	272	476	335
5	EX L	5105	4661	5679
	EX S	4168	2747	3343
	BAL	349	766	526
10	EX L	1597	3536	4284
	EX S	1697	2726	3139
	BAL	156	358	189
30	EX L	927	2739	2970
	EX S	483	1865	2173
	Variance Evaluation	12	15	15

n	Sample Type	Singh (23)	Singh (25)	Singh (27)
	BAL	614	333	812
5	EX L	4080	1931	8390
	EX S	2111	1026	4208
	BAL	1045	504	1037
10	EX L	2822	1286	6156
	EX S	2371	1174	4127
	BAL	527	173	543
30	EX L	2511	969	4742
	EX S	1547	647	2956
	Variance Evaluation	15	13	15

Bibliography

1. Cochran, W. G. (1963). Sampling Techniques. New York: John Wiley and Sons, Inc.
2. Horwitz, D. G. and Thomson, D. J. (1952). A Generalization of Sampling without Replaceemnt from a Finite Universe. A. Amer. Stat. Assn. 47, 663-685.
3. Olkin, I. (1958). Multivariate Ratio Estimation for Finite Populations. Biometrika 45, 154-155.
4. Royal, R. (1970). On Finite Populations Sampling Under Certain Linear Regression Models. Biometrika 57, 377-387.
5. Royal, R. and Herson, J. (1973). Robust Estimation in Finite Populations I. Paper accepted for publication in J. Amer. Stat. Assn.
6. Singh, M. P. (1967). Ratio Cum Product Method of Estimation. Motrika 12 (no. 1) 34-42.
7. Srivastava, S. K. (1971). Generalized Estimator for Mean of a Finite Population Using Multiple Auxiliary Information. J. Amer. Stat. Assn. 55, 404-407.
8. Greenstreet, R. L. (1973) Unpublished Ph.D. Thesis, John Hopkins School of Public Health and Hygiene, Baltimore, Md.

A REGRESSION-TYPE ESTIMATOR BASED ON PRELIMINARY TEST OF SIGNIFICANCE

J. E. Grimes and B. V. Sukhatme
California Polytechnic State University and Iowa State University

1. Introduction. If data on an auxiliary characteristic X correlated with the characteristic Y under study is available, then it is customary to use this data to provide a more efficient estimate of $\bar{Y}$, the population mean. If Y and X are correlated and the relationship between the two variables is linear, but the relationship does not pass through the origin or the correlation between Y and X is not sufficiently high, quite often a regression type estimator is used. A frequently used estimator of this type is the so-called difference estimator suggested by Hansen, Horwitz and Madow (1953), defined as

$$\bar{y}_d = \bar{y} + \beta_0(\bar{X}-\bar{x}), \tag{1.1}$$

where β_0 is a fixed constant, assumed to be known, $\bar{x}$ and $\bar{y}$ are the mean per unit estimates of $\bar{X}$ and $\bar{Y}$, and $\bar{X}$ is the population mean of X. The value of β_0 that minimizes $V(\bar{y}_d)$ is easily seen to be $\beta_2 = \sigma_{12}/\sigma_1^2$, the regression coefficient of Y on X. If no reliable guess can be made about the value of the regression coefficient, the usual practice is to estimate it from the sample by

$$\hat{\beta}_2 = s_{12}/s_1^2 \tag{1.2}$$

where $s_{12} = \sum_i^n (x_i-\bar{x})(y_i-\bar{y})/(n-1)$,

and $s_1^2 = \sum_i^n (x_i-\bar{x})^2 / (n-1)$.

and use as an estimator of $\bar{Y}$, the regression estimator $\bar{y}_\ell$ defined as,

$$\bar{y}_\ell = \bar{y} + \hat{\beta}_2(\bar{X}-\bar{x}) \ . \tag{1.3}$$

The difference estimator $\bar{y}_d$ is an unbiased estimator of the population mean $\bar{Y}$ and its variance is given by,

$$V(\bar{y}_d) = \sigma_2^2(1-\rho^2)(1+\delta^2)/n \tag{1.4}$$

where σ_1^2 and σ_2^2 are the variances of X and Y, σ_{12} is the covariance between X and Y, ρ is the correlation coefficient between X and Y and

$$\delta = (\rho - \frac{\beta_0\sigma_1}{\sigma_2}) / (1-\rho^2)^{1/2} \ . \tag{1.5}$$

The regression estimator on the other hand is generally biased, the bias vanishing when the relationship between Y and X is linear. Further its variance to terms of order n^{-2} is given by

$$V(\bar{y}_\ell) = \sigma_2^2(1-\rho^2)(n-2)/n(n-3) \ . \tag{1.6}$$

From past experience, we are often able to make an intelligent guess about β_2 the regression coefficient of Y on X. Let β_0 denote the guessed value of β_2. If β_0 is relatively close to β_2, it would appear from the above that $\bar{y}_d$ is more appropriate than $\bar{y}_\ell$ as an estimator of $\bar{Y}$, otherwise $\bar{y}_\ell$ would be more appropriate. We therefore propose an estimator which chooses between $\bar{y}_\ell$ and $\bar{y}_d$, based on a preliminary test of significance of the relative closeness of β_0 to β_2 and investigate its efficiency with respect to other regression-type estimators currently in use.

2. Proposed Regression-Type Estimator. A common method of making a test of the relative closeness of β_2 to β_0 is the usage of the statistic,

$$t = \sqrt{n-2}(\hat{\beta}_2-\beta_0)s_1 / (s_2^2 - \hat{\beta}_2^2 s_1^2)^{1/2} \tag{2.1}$$

where $s_2^2 = \sum_i^n (y_i-\bar{y})^2 / (n-1)$. (2.2)

If from past experience, it is hypothesized that β_2 is β_0 but nothing further is known about β_2, the proposed estimator based on preliminary test of significance, to be called Sometimes Regression Estimator, may be defined as

$$\begin{aligned} \bar{y}_s &= y_d \quad \text{if} \quad t \in A \\ &= \bar{y}_\ell \quad \text{if} \quad t \in A^c \end{aligned} \tag{2.3}$$

where A is the event $|t| \le t_0$ and A^c the complementary event $|t| > t_0$.

Now we need to look at a criterion for deciding whether or not the proposed estimator $\bar{y}_s$ has any advantages over $\bar{y}_d$ and $\bar{y}_\ell$. The most commonly used loss function is the squared error. This then leads to considering the variance of the estimator $\bar{y}_s$ if it is unbiased, or the mean square error of $\bar{y}_s$ if it is biased. We then have the expected value of $\bar{y}_s$ given by

$$E(\bar{y}_s) = E(\bar{y}_d|A)P(A) + E(\bar{y}_\ell|A^c)P(A^c), \tag{2.4}$$

and the mean square error of $\bar{y}_s$ is given by

$$\begin{aligned} M.S.E.(\bar{y}_s) = E(\bar{y}_s-\bar{Y})^2 &= E[(\bar{y}_d-\bar{Y})^2|A]P(A) \\ &+ E[(\bar{y}_\ell-\bar{Y})^2|A^c]P(A^c) \ . \end{aligned} \tag{2.5}$$

3. Expected Value and Variance of $\bar{y}_s$. It is necessary to make suitable assumptions about the joint distribution of X and Y in order to obtain a closed form for the expected value and the variance of $\bar{y}_s$. In what follows, we assume that the population is infinite and that X and Y have a bivariate normal distribution function.

Theorem 3.1: $\bar{y}_s$ is an unbiased estimator of the population mean $\bar{Y}$.

Proof: Using the fact that $\bar{x}$ and (s_1^2, s_2^2, s_{12}) are statistically independent, it can be easily seen that $E(\bar{y}_s) = \bar{Y}$. Q.E.D.

Since $\bar{y}_s$ is an unbiased estimator, we now obtain the variance of $\bar{y}_s$. As $(\bar{x},\bar{y})$ and (s_1^2, s_2^2, s_{12}) are statistically independent, we have from (2.5)

$$V(\bar{y}_s) = V(\bar{y}_d) - \frac{2\sigma_{12}}{n} E[(\hat{\beta}_2 - \beta_0)|A^c]P(A^c) + \frac{2\beta_0\sigma_1^2}{n} E[(\hat{\beta}_2 - \beta_0)|A^c]P(A^c) + \frac{\sigma_1^2}{n} E[(\hat{\beta}_2 - \beta_0)^2|A^c]P(A^c) \quad . \qquad (3.1)$$

In order to further evaluate this, we need an expression for $E[(\hat{\beta}_2 - \beta_0)^h|A^c]P(A^c)$ for $h = 0,1,2$. It will be assumed that the sample size is $n \geq 4$.

Lemma 3.2: $KP(|t|>t_0)E[(\hat{\beta}_2 - \beta_0)^h||t|>t_0]$

$$= \sum_{i=0}^{\infty} (2\theta)^{2i} \frac{\Gamma(\frac{h+2i+1}{2})\Gamma(\frac{n+2i-h-1}{2})}{\Gamma(2i+1)} I(h+2i+1) \text{ if } h \text{ is even,}$$

$$= \sum_{i=0}^{\infty} (2\theta)^{2i+1} \frac{\Gamma(\frac{h+2i+2}{2})\Gamma(\frac{n+2i-h}{2})}{\Gamma(2i+2)} I(h+2i+2), \text{ if } h$$

is odd where $m_0 = (n-2)/[t_0^2 + (n-2)]$,

$$K = \sqrt{\pi}\,\Gamma(\tfrac{n-1}{2})(1+\delta^2)^{\frac{n-h-1}{2}} (\sigma_1/\sigma_2\sqrt{1-\rho^2})^h, \quad \theta = \frac{\delta}{\sqrt{1+\delta^2}},$$

$I.(\cdot,\cdot)$ is the incomplete beta distribution function and $I(x)$ denotes $I_{m_0}(\frac{n-2}{2}, \frac{x}{2})$.

Proof: It is well-known that the joint density function for s_1, s_2 and $r = s_{12}/s_1 s_2$ is given by

$$f(s_1,s_2,r) = K_1 (s_1^2 s_2^2)^{\frac{n-2}{2}} (1-r^2)^{\frac{n-4}{2}} \times \exp[-\frac{n-1}{2(1-\rho^2)}(\frac{s_1^2}{\sigma_1^2} - \frac{2\rho s_1 s_2 r}{\sigma_1\sigma_2} + \frac{s_2^2}{\sigma_2^2})]$$

if $0 < s_1 < \infty$, $0 < s_2 < \infty$, and $r^2 < 1$,

$= 0$ otherwise,

where $K_1 = (n-1)^{n-1}/\pi\Gamma(n-2)[(1-\rho^2)\sigma_1^2\sigma_2^2]^{\frac{n-1}{2}}$.

Making the transformation

$u = (n-1)s_1^2/2\sigma_1^2(1-\rho^2)$, $v = (n-1)r s_1 s_2/2\sigma_1\sigma_2(1-\rho^2)$ and

$t' = t/\sqrt{n-2} = (\hat{\beta}_2 - \beta_0)s_1/(s_2^2 - \hat{\beta}_2^2 s_1^2)^{1/2}$, we get

$$f(u,v,t') = \frac{K_3}{ut'^{n-1}} \exp[-u(1-\rho^2)(1+\delta^2) - \frac{1+t'^2}{ut'^2}(v - \frac{\beta_0 u\sigma_1}{\sigma_2})^2] \times \sum_{i=0}^{\infty} \frac{2^i (v - \frac{\beta_0 u\sigma_1}{\sigma_2})^{n+i-2} \cdot \delta^i (1-\rho^2)^{\frac{i}{2}}}{\Gamma(i+1)}$$

in $R_1 = (0 \leq u < \infty,\ 0 \leq t' < \infty,\ v \geq u\sigma_1\beta_0/\sigma_2)$

$$= \frac{K_3}{u|t'|^{n-1}} \exp[-u(1-\rho^2)(1+\delta^2) - \frac{1+t'^2}{ut'^2}(v - \frac{\beta_0 u\sigma_1}{\sigma_2})^2] \times \sum_{i=0}^{\infty} \frac{(-1)^i 2^i \left|v - \frac{\beta_0 u\sigma_1}{\sigma_2}\right|^{n+i-2} \delta^i (1-\rho^2)^{\frac{i}{2}}}{\Gamma(i+1)},$$

in $R_2 = (0 \leq u < \infty,\ -\infty < t' < 0,\ v < u\sigma_1\beta_0/\sigma_2)$

$= 0$, otherwise,

where $K_3 = 2^{n-2}(1-\rho^2)^{\frac{n-1}{2}} / \pi\Gamma(n-2)$.

We have $P(|t'|>t_0')E((\hat{\beta}_2 - \beta_0)^h \big| |t'| > t_0'$

$$= \int_{R_4} [(v - \frac{\beta_0 u\sigma_1}{\sigma_2})\frac{\sigma_2}{u\sigma_1}]^h f(u,v,t')\,du\,dv\,dt' + \int_{R_5} [(v - \frac{\beta_0 u\sigma_1}{\sigma_2})\frac{\sigma_2}{u\sigma_1}]^h f(u,v,t')\,du\,dv\,dt'$$

$$= I_4 + I_5$$

where $R_4 = \{0 < u < \infty,\ -\infty < t' < t_0',\ v < u\sigma_1\beta_0/\sigma_2\}$, and

$R_5 = \{0 \leq u \leq \infty,\ v \geq u\sigma_1\beta_0/\sigma_2,\ t_0' < t' < \infty\}$.

To obtain the desired result the following lemmas are needed.

Lemma 3.3: $\int_{-\infty}^{0} |x|^n e^{-\frac{1}{2}x^2} = 2^{\frac{n-1}{2}}\Gamma(\frac{n+1}{2})$.

Lemma 3.4: $\Gamma(\frac{j-2}{2})\,\Gamma(\frac{j-1}{2})\,2^{j-3} = \Gamma(j-2)\sqrt{\pi}$.

Now, $I_4 = K_3 \int_{-\infty}^{t_0'} \int_{0}^{\infty} \int_{-\infty}^{u\beta_0\sigma_1/\sigma_2} (\frac{2}{\sigma_1})^h (\frac{1}{u})^{h+1} \frac{1}{|t'|^{n-1}}$

$$x \exp[-u(1-\rho^2)(1+\delta^2) - \frac{1+t'^2}{ut'^2}(v-\beta)^2]$$

$$\sum_{i+0}^{\infty} (-1)^{h+i} \frac{\left|v - \frac{\beta_0 u\sigma_1}{\sigma_2}\right|^{n+h+i-2} (2\delta\sqrt{1-\rho^2})^i dv\, du\, dt'}{\Gamma(i+1)}$$

$$= \frac{1}{2K} \sum_{i=0}^{\infty} (-1)^{h+i} (2\theta)^i \frac{\Gamma(\frac{h+i+1}{2})\Gamma(\frac{n+i-h-1}{2})}{\Gamma(i+1)} I(h+i+1).$$

Similarly I_5 can be obtained. Q.E.D.

Using Lemma 3.2 and substituting into (3.1) we obtain the following theorem.

Theorem 3.5: $V(\bar{y}_s) - V(\bar{y}_d)$

$$= \frac{2\sigma_2^2(1-\rho^2)}{n} \sum_{i=0}^{\infty} \frac{\Gamma(\frac{n+2i-1}{2})\delta^{2i+2}}{\Gamma(i+1)\Gamma(\frac{n-1}{2})(1+\delta^2)^{\frac{n+2i-1}{2}}} I(2i+3)$$

$$+ \frac{\sigma_2^2(1-\rho^2)}{n} \sum_{i=0}^{\infty} \frac{(2i+1)\Gamma(\frac{n+2i-3}{2})\delta^{2i}}{2\Gamma(i+1)\Gamma(\frac{n-1}{2})(1+\delta^2)^{\frac{n+2i-3}{2}}} I(2i+3) .$$

As t_0 tends to infinity, $V(\bar{y}_s)$ tends to $V(\bar{y}_d)$ as is to be expected since the estimator $\bar{y}_s$ becomes $\bar{y}_d$. Similarly, as t_0 tends to zero, $V(\bar{y}_s)$ tends to $V(\bar{y}_\ell)$ since the estimator $\bar{y}_s$ becomes $\bar{y}_\ell$.

4. Comparison of Different Estimators

A. Comparison of the sometimes regression estimator with the difference estimator.

Consider

$$D_2(\theta, m_0) = n\Gamma(\tfrac{n-1}{2})(1+\delta^2)^{\frac{n-3}{2}}(V(\bar{y}_s) - V(\bar{y}_d))/\sigma_2^2(1-\rho^2) \tag{4.1}$$

Then, we have from Theorem 3.5

$$D_2(\theta, m_0) = \sum_{j=0}^{\infty} \frac{\Gamma(\frac{n+2j-3}{2})\theta^{2j}}{2\Gamma(j+1)} I(2j+3)[(2j+1) - 2(n+2j-3)\theta^2] . \tag{4.2}$$

Let j=i-1 in the first summation of (4.2), then we have

$$D_2(\theta, m_0) = \tfrac{1}{2}\Gamma(\tfrac{n-3}{2})I(3) + 2\sum_{j=0}^{\infty} \frac{\Gamma(\frac{n+2j-1}{2})\theta^{2j+2}}{\Gamma(j+1)} I(2j+5)[\frac{2j+3}{4(j+1)} - \frac{I(2j+3)}{I(2j+5)}] \tag{4.3}$$

$$= \sum_{j=0}^{\infty} C_j(m_0)\theta^{2j} , \tag{4.4}$$

where $C_0(m_0) = \frac{1}{2}\Gamma(\frac{n-3}{2})I_{m_0}(\frac{n-2}{2}, \frac{3}{2})$ (4.5)

and $C_{j+1}(m_0) = \frac{2\Gamma(\frac{n+2j-1}{2})I(2j+5)}{\Gamma(j+1)}$

$$[\frac{2j+3}{4(j+1)} - \frac{I(2j+3)}{I(2j+5)}], \quad j=0,1,2,\ldots . \tag{4.6}$$

Consider first the effect of variation in θ. θ will vary over the interval (-1,1) since δ may vary over the interval $(-\infty, \infty)$.

Lemma 4.1: For a = 1, $\frac{3}{2}$, 2, $\frac{5}{2}$ ... and

c = 1, $\frac{3}{2}$, 2, $\frac{5}{2}$, ...

$$\frac{I_x(a,c)}{I_x(a,c+1)} \Big/ \frac{I_x(a,c+\frac{1}{2})}{I_x(a,c+\frac{3}{2})} \leq 1, \quad \text{for} \quad 0 < x \leq 1 .$$

Proof: L'Hospital's rule may be used to show that the lemma holds in a positive neighborhood of zero. Then the lemma may be proved for the entire interval by defining $\phi(x) = I_x(a, c+1)$. $I_x(a,c+\frac{1}{2}) - I_x(a,c)I_x(a,c+\frac{3}{2})$, and showing that there exists an x_1 such that

$$\phi'(x) \geq 0 \qquad 0 < x \leq x_1 ,$$
$$< 0 \qquad x_1 < x \leq 1 .$$

Lemma 4.2: For $0 < m_0 \leq 1$ $C_0(m_0)$, $C_1(m_0)$, $C_2(m_0)$, ... is a sequence of numbers such that for some $K > 0$

$$C_j(m_0) \geq 0 \qquad j \leq K$$
$$< 0 \qquad j > K .$$

Proof: Since $(2j+3)/4(j+1)$ is a decreasing function of j, and by Lemma 4.1, $I(2j+3)/I(2j+5)$ $j = 0,1,2,\ldots$ is an increasing function of j, $(2j+3)/4(j+1) - I(2j+3)/I(2j+5)$ $j = 0,1,2,\ldots$ is a decreasing function of j. Now, $C_0(m_0) > 0$. Suppose that $C_j(m_0) \geq 0$ for $j = 1,2,\ldots$. This implies that $D_2(\theta, m_0) \geq 0$ for $0 \leq \theta < 1$. But from (4.2) for $\theta = 1/\sqrt{2} + \varepsilon$ with $\varepsilon > 0$, $D_2(1/\sqrt{2} + \varepsilon, m_0) < 0$ for $0 < m_0 \leq 1$ leading to contradiction. Hence the lemma is proven. Q.E.D.

Define the relative efficiency of $\bar{y}_s$ with respect to $\bar{y}_d$ as $e_2(\delta, m_0) = V(\bar{y}_d)/V(\bar{y}_s)$.

Theorem 4.3: For m_0 fixed such that $0 < m_0 \leq 1$, there exists a θ_0 where $0 < \theta_0 < 1$ and

$D_2(\theta, m_0) > 0$ and hence $e_2(\delta, m_0) < 1$, $-\theta_0 < \theta < \theta_0$

≤ 0 and hence $e_2(\delta, m_0) \geq 1$ otherwise.

Proof: Since from (4.3) $D_2(\theta, m_0)$ is symmetric

in θ, it is necessary only to consider $D_2(\theta, m_0)$ for θ positive.

From (4.2), $D_2(0, m_0) > 0$ for $0 < m_0 \leq 1$.

Further for $\theta = 1\sqrt{2} + \epsilon$ with $\epsilon > 0$,
$[(2j+1) - 2(n+2j-3)\theta^2] < 0 \quad j = 0,1,2,\ldots,$
and we have $D_2(1/\sqrt{2} + \epsilon, m_0) < 0$, $0 < m_0 \leq 1$.
Since $D_2(\theta, m_0)$ is continuous then there exists θ_0 such that $0 < \theta_0 < 1$ and $D_2(\theta_0, m_0) = 0$.

We now show that $D_2(\theta, m_0) < 0$ for $\theta > \theta_0$. By Lemma 4.2 there exists a K such that

$$C_j(m_0) > 0 \quad \text{for } j < K$$
$$\leq 0 \quad \text{for } j \geq K .$$

Hence $\sum_{j=0}^{\infty} C_j(m_0)\theta_0^{2j} = 0$ i.e., $\sum_{j=0}^{\infty} C_j(m_0)\theta_0^{2j-1} = 0$;

and since $D_2(\theta, m_0)$ is a power series in θ which converges for $-1 < \theta < 1$, we get

$$\frac{\partial D_2(\theta, m_0)}{\partial \theta} = \sum_{j=0}^{\infty} 2j\, C_j(m_0)\theta^{2j-1} \quad \text{for } 0 \leq \theta < 1 ,$$

and therefore $\left.\frac{\partial D_2(\theta, m_0)}{\partial \theta}\right|_{\theta_0}$

$$\leq 2K \sum_{j=1}^{\infty} C_j(m_0)\theta_0^{2j-1} = \frac{-2K\, C_0(m_0)}{\theta_0} < 0 .$$

It can be similarly shown that if $D_2(\theta^*, m_0) < 0$, then $\left.\frac{\partial D_2(\theta, m_0)}{\partial \theta}\right|_{\theta^*} < 0$. Therefore for m_0 fixed, as θ increases, $D_2(\theta, m_0)$ becomes negative and stays negative. Q.E.D.

Next consider the variation of $D_2(\theta, m_0)$ due to m_0 with θ fixed.

Lemma 4.4: If for fixed θ, there exists an $m_0^* \in (0,1)$ such that

$$\left.\frac{\partial D_2(\theta, m_0)}{\partial m_0}\right|_{m^*} = 0$$

then $\frac{\partial D_2(\theta, m_0)}{\partial m_0} > 0 \quad 0 \leq m_0 < m_0^*$

$$= 0 \qquad m_0 = m_0^*$$
$$< 0 \quad m_0^* < m_0 < 1 .$$

The proof of this lemma follows in a manner analogous to the proof of Theorem 4.3.

Theorem 4.5: There exists $\theta_1^* > 0$ and $\theta_2^* > 0$ defined by $D_2(\theta_1^*, 1) = 0$, and $\theta_2^* = \inf_S \theta$, where $S = \{\theta : \theta > 0,\ D_2(\theta, m_0) \leq 0$ for all $m_0 \ni 0 < m_0 \leq 1\}$; such that

a) for θ fixed and $\epsilon[-\theta_1^*, \theta_1^*]$ $D_2(\theta, m_0) \geq 0$ and hence $e_2(\delta, m_0) \leq 1$ for $0 < m_0 < 1$,

b) for θ fixed and $\epsilon\{(-\theta_2^*, -\theta_1^*) \cup (\theta_1^*, \theta_2^*)\}$, $\exists\, m_0^* \ni 0 < m_0^* < 1$, and
$D_2(\theta, m_0) \geq 0$ and hence
$e_2(\delta, m_0) \leq 1 \qquad 0 < m_0 \leq m_0^*$,
$D_2(\theta, m_0) < 0$ and hence
$e_2(\delta, m_0) > 1 \qquad m_0^* < m_0 \leq 1$;

c) for θ fixed and $\epsilon\{(-1, -\theta_2^*] \cup [\theta_2^*, 1)\}$
$D_2(\theta, m_0) \leq 0$, and hence $e_2(\delta, m_0) \geq 1$ for $0 < m_0 \leq 1$.

Proof: Since $D_2(\theta, m_0)$ is symmetric in θ, it is necessary only to consider $D_2(\theta, m_0)$ for $\theta > 0$. Suppose for θ fixed $\ni 0 < \theta < 1$, $\exists\, m_0^* \ni m_0^* \in (0,1)$ and $D_2(\theta, m_0^*) = 0$. Since

$$\lim_{m_0 \to 0} D_2(\theta, m_0) = \lim_{m_0 \to 0} \frac{\partial D_2(\theta, m_0)}{\partial m_0} = 0,$$ it follows from Lemma 4.4 that if $\frac{\partial D_2(\theta, m_0)}{\partial m_0} < 0$ in the neighborhood of $m_0 = 0$, then $\frac{\partial D_2(\theta, m_0)}{\partial m_0} < 0$ $0 < m_0 \leq 1$. Under that condition there could not be a point $m_0^* \ni 0 < m_0^* \leq 1$ and $D_2(\theta, m_0) = 0$.

Hence in order that $D_2(\theta, m_0^*) = 0$ it follows that there must exist an m_0^{**} such that $0 < m_0^{**} < m_0^* \leq 1$ and

$$\frac{\partial D_2(\theta, m_0)}{\partial m_0} > 0 \qquad 0 < m_0 < m_0^{**}$$
$$= 0 \qquad m = m_0^{**}$$
$$< 0 \qquad m_0^{**} < m_0 \leq 1 .$$

Hence if $D_2(\theta, m_0^*) = 0$ then for $m_0 > m_0^*$, $D_2(\theta, m_0) < 0$. By above if for $\theta = \theta_1$, $D_2(\theta_1, 1) \geq 0$ then $D_2(\theta_1, m_0) \geq 0$, $0 < m_0 \leq 1$. If further for $\theta = \theta_2$, $D_2(\theta_2, 1) < 0$, then by Theorem 4.3, $\theta_2 > \theta_1$. Hence $\theta_1^* = \{\theta : \theta > 0$ and $D_2(\theta, 1) = 0\}$. If $D_2(\theta, 1) < 0$ then either $\theta = \theta_3$ and $D_2(\theta_3, m_0) \leq 0$, $0 < m_0 \leq 1$ or $\theta = \theta_4$ and

$$\exists\, m_0^* \ni D_2(\theta_4, m_0) > 0 \qquad 0 < m_0 < m_0^*$$
$$= 0 \qquad m_0 = m_0^*$$
$$< 0 \qquad m_0^* < m_0 \leq 1 .$$

Now for $m_{0_1} < m_0^*$, $D_2(\theta_3, m_{0_1}) \leq 0$ and $D_2(\theta_4, m_{0_1}) \geq 0$, then by Theorem 4.3 $\theta_4 \leq \theta_3$.

Hence $\theta_2^* = \underset{S}{\text{Inf}}\ \theta$ and theorem is proved. Q.E.D.

Theorem 4.6: For e_0 fixed such that $0 < e_0 < 1$, there exists an m_0^* such that for $m_0 \leq m_0^*$, $e_2(\delta, m_0) \geq e_0$.

Proof: By Lemma 4.4, for fixed θ or equivalently for fixed δ, $\exists\, m_0(\theta) \ni$

$$e_2(\delta, m_0) = 1 \Big/ \left[1 + \frac{D_2(\theta, m_0)}{\Gamma(\frac{n-1}{2})(1+\delta^2)^{\frac{n-1}{2}}}\right] \geq e_0 \quad \text{for}$$

$0 < m_0 \leq m_0(\theta)$. Pick $m_0^* = \underset{0 \leq \theta < 1}{\text{Inf}}\ m_0(\theta)$. Hence

$e_2(\delta, m_0) \geq e_2(m_0^*, \delta) \geq e_0$ for $0 < m_0 \leq m_0^*$ and for any $\delta \in [0, \infty)$. Q.E.D.

B. Comparison of the Sometimes Regression Estimator with the Regression Estimator.

$$\text{Let } D_1(\theta, m_0) = n(V(\bar{y}_s) - V(\bar{y}_\ell)) / \sigma_2^2(1-\rho^2)(1-\theta^2)^{\frac{n-3}{2}} \tag{4.7}$$

Then using (1.4), (1.6), (4.1) and (4.3), it can be seen that

$$D_1(\theta, m_0) = \theta^2(1-\theta^2)^{-\frac{n-1}{2}} - \frac{(1-\theta^2)^{-\frac{n-3}{2}}}{n-3} + I(3)/(n-3)$$
$$+ \sum_{j=0}^{\infty} \frac{2\Gamma(\frac{n+2j-1}{2})\theta^{2j+2}}{\Gamma(j+1)\Gamma(\frac{n-1}{2})} I(2j+5)$$
$$\times \left[\frac{2j+3}{4(j+1)} - \frac{I(2j+3)}{I(2j+5)}\right] . \tag{4.8}$$

Define the relative efficiency of $\bar{y}_s$ with respect to $\bar{y}_\ell$ as $e_1(\delta, m_0) = V(\bar{y}_\ell) / V(\bar{y}_s)$. Consider first the effect of variation of θ.

Theorem 4.7: For m_0 fixed such that $0 < m_0 \leq 1$, there exists a θ_0 such that $0 < \theta_0 < 1$ and $D_1(\theta, m_0) < 0$ and hence $e_1(\delta, m_0) > 1$, $-\theta_0 < \theta < \theta_0$

≥ 0 and hence $e_1(\delta, m_0) \leq 1$ otherwise.

The theorem can be proved by using techniques similar to those used in proving Theorem 4.3.

Next consider the effect of m_0 with θ fixed. The result is given without proof in Theorem 4.8.

Theorem 4.8: With θ fixed, $D_1(\theta, m_0)$ varies with $D_2(\theta, m_0)$ as a function of m_0. For θ fixed such that $0 \leq \theta < 1$, $D_1(\theta, m_0)$ falls in one of the following three categories:

(a) $D_1(\theta, m_0)$ is always increasing as a function of m_0 for $0 < m_0 \leq 1$;

(b) $\exists\, m_0^*$ such that $0 < m_0^* < 1$ and $D_1(\theta, m_0)$ is increasing as a function of m_0 for $m_0 < m_0^*$ and decreasing for $m_0 > m_0^*$;

(c) $D_1(\theta, m_0)$ is always decreasing as a function of m_0 for $0 < m_0 \leq 1$.

5. Conclusions and Recommendations Regarding the Use of the Sometimes Regression Estimator.

If conditions are such that the use of regression type estimators is warranted, the question arises as to when the sometimes regression estimator would be most appropriate. Actually, the sometimes regression estimator includes both the difference estimator $\bar{y}_d$ and the regression estimator $\bar{y}_\ell$ as special cases. Hence the sometimes regression estimator may be used whenever it is appropriate to use regression type estimators.

Consider the effect of change in the relative closeness of β_0 to β_2. Theorem 4.3 gives the result that for fixed m_0, $V(\bar{y}_s)$ is greater than $V(\bar{y}_d)$ for β_0 close to β_2, but this relationship reverses itself as the distance of β_2 from β_0 increases and it remains reversed. Theorem 4.7 illustrates that the situation is reversed for the relationship of the variance of the sometimes regression estimator to the variance of the regression estimator. Analogous results hold for the relative efficiencies. These results are illustrated in Figures 1 and 2 for n equal to 6. The relative distance between β_2 and β_0 is a fixed unknown quantity. However on the basis of past experience, it may be possible to have some idea about the likely range of values it can take on.

Now m_0 can be fixed in any manner we please. If m_0 is fixed such that the probability of using $\bar{y}_d$ is very high, then the relative efficiency of $\bar{y}_s$ with respect to $\bar{y}_d$ is close to 1. On the other hand, if m_0 is such that the probability of using $\bar{y}_\ell$ is high, then the relative efficiency of $\bar{y}_s$ with respect to $\bar{y}_\ell$ is close to 1. The effect of changing the level of significance of the test when the relative distance between β_0 and β_2 is

fixed is illustrated in Figures 3 and 4 for n equal to 6.

If there is a priori information that β_0 may be the actual value of β_2, the guidelines for using the sometimes regression estimator may be stated as follows:

1) If β_0 is considered a very reliable guess for β_2 then t_0 may be chosen so that the likelihood that $\bar{y}_s$ results in using $\bar{y}_d$ is high. This would tend to minimize the loss in efficiency of $\bar{y}_s$ with respect to $\bar{y}_d$.

2) If β_0 is not considered a very reliable choice for β_2 then t_0 may be chosen so that the likelihood that $\bar{y}_s$ results in using $\bar{y}_\ell$ is very high. This would tend to minimize the loss in efficiency of $\bar{y}_s$ with respect to $\bar{y}_\ell$.

3) If no further information is available about the reliability of the choice of β_0, a middle range value for m_0 may be used.

6. Acknowledgement. This work was supported under Contract No. OEC-0-73-6640 by the U.S. Office of Education, Department of Health, Education and Welfare.

REFERENCES

Cochran, W.G. 1963. Sampling Techniques, Second Edition, Wiley, New York.

Cramer, Harold, 1946. Mathematical Methods of Statistics. Princeton, N.J., Princeton University Press.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. 1953. Sample survey methods and theory Vols. I and II: Methods and applications. Wiley, N.Y.

Sukhatme, P.V. and Sukhatme, B.V. 1970. Sampling theory of surveys with applications. Ames, Iowa, Iowa State University Press.

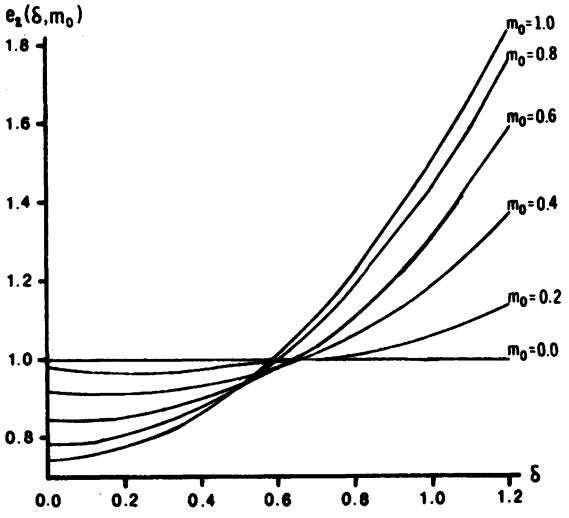


FIGURE 1

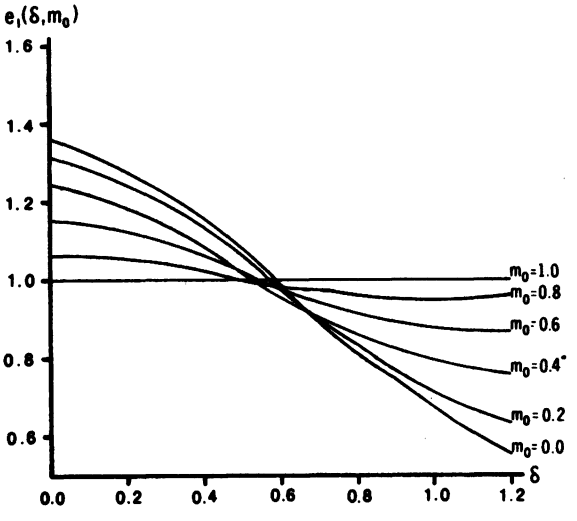


FIGURE 2

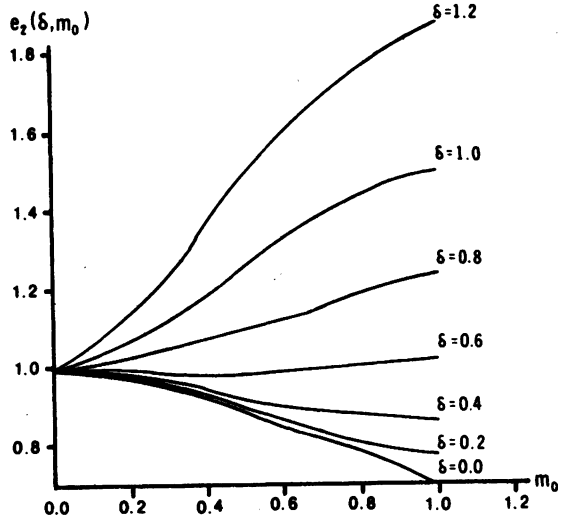


FIGURE 3

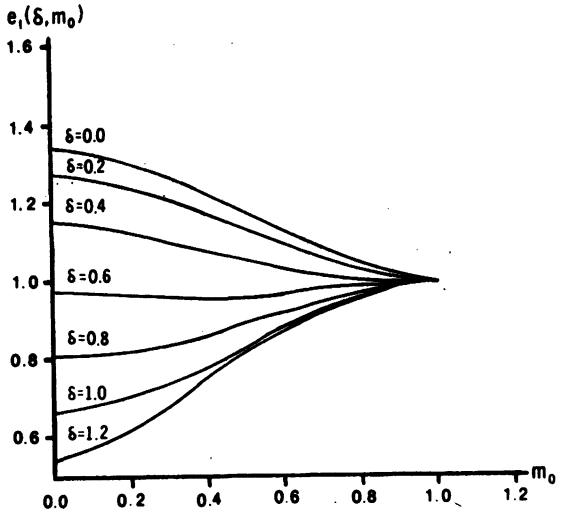


FIGURE 4

A MODEL OF FERTILITY BEHAVIOUR IN INDIA[1]

Kanhaya L. Gupta, University of Alberta

Recently, Professor Bhagwati wrote that, "For a number of underdeveloped countries, there is an important choice in the use of resources for reducing the growth of population. Indeed, it is vitally necessary for countries with severe underemployment and rapidly growing populations to consider the question of population control with utmost seriousness. Unfortunately, even countries in serious population difficulties(as, for example, India) have contented themselves with cursory analysis and inadequate action in this important field."[2]

This paper is an attempt to contribute to our understanding of the factors which determine the birth rate in India. In this paper I specify and estimate a simultaneous equations model of fertility behaviour in India. The model has five equations and five endogenous variables. The endogenous variables are: birth rate, infant mortality, per capita income, dependency rate, and female participation rate. This specification allows for interdependence between birth rate and per capita income as suggested by Okun(1965); takes account of the opportunity cost of child bearing as suggested by Mincer (1963), Cain(1966), and Schultz(1969)and treats a wife's labour force participation decision as endogenous as suggested by Willis(1973).

The scheme of the paper is as follows. In Section I, I specify and discuss the model. In Section II, I discuss the data and present the results. This is followed by a discussion of the results and the multiplier analysis in Section III. The last section briefly summarizes the main findings.

I. THE MODEL

The equations of the model are as follows:

1) $BR=a_1+\overset{+}{b}_1 IMR+\overset{+}{c}_1 Y+\overset{-}{d}_1 LR+\overset{-}{e}_1 FPR+\overset{-}{f}_1 AMF$

2) $IMR=a_2+\overset{-}{b}_2 Y+\overset{-}{c}_3 LR+\overset{-}{d}_4 HE$

3) $Y=a_3+\overset{-}{b}_3 ZD+\overset{+}{c}_3 LR+\overset{+}{d}_3 KP+\overset{-}{e}_3 ZT$

4) $FPR=a_4+\overset{-}{b}_4 BR+\overset{+}{c}_4 LA+\overset{+}{d}_4 FLR$

5) $ZT=a_5+\overset{+}{b}_5 BR$

where BR* stands for birth rate, IMR* for infant mortality rate, Y* for per capita income, FPR* for female participation rate, LR for literacy rate, AMF for age at marriage of females in urban areas, HE for per capita expenditure on health services, ZD for population density, KP for per capita energy consumption, FLR for female literacy rate, LA for percentage of male labour force in agriculture, and ZT* for percentage of population below 15 years of age(dependency rate). The variables with an asterisk are endogenous while the remaining ones are endogenous.[3] The expected signs of the parameters are shown above the parameters in each equation.

Some comments are in order on these equations.

Fertility equation

This equation embodies the hypotheses put forward by Adelman(1963), Becker (1960), Mincer(1963), Cain(1966), and Schultz(1969,1973), among others. Very briefly, the inclusion of per capita income follows from the theory of consumer choice, as for example, argued by Becker. Female participation rate is included as a proxy for the, "opportunity income of women and their access to the labor market".[4] The inclusion of infant mortality is justified in terms of the replacement needs of a family for children [Gregory et al.(1972), Schultz(1973)]. The role of education has been discussed extensively in the literature and does not need any elaboration. The reason for the inclusion of age at marriage is that women marrying at an earlier age are exposed to sexual activity for a longer duration and thus the possibility of a larger number of births. Thus we would expect a negative relationship, cateris paribus, between fertility rate and the age at marriage (Driver 1963). A specific point is in order about the particular variable used in this study. The proper variable to use would have been the age at marriage of females in both the urban and the rural areas. But unfortunately the only information currently available relates to the age at marriage of females in the urban areas only. Since, however, most of India's population is still rural, where the age at marriage of females is somewhat lower, our measure most likely overestimates the actual age.

Infant mortality equation

This equation is quite similar to the one used by Adelman. The variable, per capita health expenditure, is used as an index of the availability of health care services.

Per capita income equation

In this equation, per capita energy consumption is used as a proxy for per capita fixed capital.[5] Population density is used to measure the pressure of population on non-capital resources, although this is only a rough measure as pointed out by Adelman. Literacy may be taken to represent quality of labour. As already pointed out, this model allows for interdependence between the birth rate and per capita income according to Okun's arguments. We hypothesize that birth rate affects per capita income in the following way.[6] Per capita income, cateris paribus, depends on total labour force participation rate. Total labour force participation rate, among other things,

depends on dependency rate. And as we shall argue below, dependency rate depends on birth rates. Thus our hypothesis is that high birth rates lower per capita income by leading to high dependency rate which in its turn leads to lower total labor force participation rates. A priori, therefore, we would expect a negative partial relationship between per capita income and dependency rates.[7]

Female participation rate

Birth rate affects the supply of female labour and is therefore included as an argument of the female participation rate equation.[8] It is now well recognized that labour force participation is higher in largely agrarian economies than in industrialized economies,[9] and we therefore expect a positive relationship between female participation rate and the level of non-industrial development where the latter is measured by the percentage of male labour force in agriculture. We expect the effect of female literacy rate to be positive because more jobs are open to educated women than to the uneducated.[10]

Dependency rate equation

This equation is straightforward and its rationale based on demographic theory has been ably summarized by Leff. Thus, "demographic theory indicates that a prolonged high birth rate will affect a population's age composition, placing a relatively large percentage of population in the younger age bracket".[11]

II. THE DATA AND THE RESULTS

The model was estimated by the method of Two Stage Least Squares using cross-section data for 1961 for thirteen states.[12] These were the only data that could be found for such a study.[13] The details are given in the Appendix. The 't' values are given in the parenthesis.

The results are given below.

1) $BR = 21.6412 + 0.1196IMR + 0.0364y - 0.0701LR$
(1.818) (1.249) (0.143)
$-0.0041FPR - 0.3590AMF$
(0.030) (0.065) $R^2 = 0.2213$

2) $IMR = 237.0754 - 0.2758y - 0.1284LR - 2.9850HE$
(1.558) (0.093) (0.162)
$R^2 = 0.1410$

3) $y = 108.4419 - 0.0297ZD + 5.5328LR + 0.0824KP$
(0.323) (1.621) (1.658)
$-2.1068ZT$
(0.179) $R^2 = 0.3940$

4) $FPR = -20.8348 - 1.9762BR + 1.6891LA$
(1.713) (2.361)
$+1.1258FLR$
(1.470) $R^2 = 0.4310$

5) $ZT = 16.7043 + 0.5580BR$
(7.966) $R^2 = 0.8710$

III. DISCUSSION OF THE RESULTS

In so far as the signs of different coefficients are concerned, the model must be considered a success. However in terms of the statistical significance of the coefficients and the values of R^2 the results are somewhat mixed.

In the fertility equation, only the coefficients of infant mortality and per capita income exceed their standard errors. Despite the fact that the explanatory power of this equation is low, the results are nevertheless highly suggestive. For one thing our results support the finding of a significant partial relationship between fertility and infant mortality for a large number of low income countries[14] and thus provides additional evidence for Schultz's contention that in our efforts to explain fertility behaviour, "the regime of mortality cannot be neglected in low-income countries."[15] Our results, based on highly aggregative data, which however cover the entire country, also provide support for Driver's(1963) finding of a negative relationship between age at marriage and fertility. It should be pointed out here that Driver's study was based on a sample of about 2600 households from Central India. Our results also accord with those of Gregory et al.[16] who in a two equation model of fifteen developing countries found female participation rate and literacy to be insignificant variables in their birth rate equation.

In the infant mortality equation, only the coefficient of per capita income exceeds its standard error. The other two variables are only suggestive.

The results of the per capita income equation are fairly reasonable. The unimportance of population density has also been reported by Gregory et al.(1972). The lack of significance of the dependency variable deserves special mention. In a slightly different context -- dependency rates as being a determinant of savings rate in LDC's -- I reported earlier that in very poor countries, the relation between these variables was insignificant.[17] The present result would thus appear to be consistent with that finding. An interesting thing about this simultaneous equations model should be pointed out here. That relates to the role of literacy. In single equation models, literacy influences fertility only directly as in our equation (1). However, in our model, it affects indirectly too -- through its effects on infant mortality, per capita income and female participation rate in so far as female literacy is concerned. Thus it would appear that single equation models seriously underestimate the role of education in dealing with population control.

The equations for female participation rates and dependency rate are highly satisfactory. Every coefficient exceeds its standard error.

In order to examine the impact of exogenous variables, we solved the model for its reduced form. However, since the units

of measurement are not common for all the variables, to make the reduced form coefficients comparable, using means of the variables, I converted them into elasticities. These elasticity multipliers measure the direct and indirect effects of a one per cent change in a given exogenous variable on a given endogenous variable. The results are given in Table 1. (See the Appendix for the tables).

Looking at the elasticity multipliers of the birth rate equation, the results are very disheartening. Not one of them amounts to anything. Considering literacy rate, for example, a major policy variable, we see that a one hundred per cent increase in literacy will cause a mere three per cent reduction in the birth rate. This means that even if India could achieve complete literacy -- from the present level of thirty per cent -- its birth rate will decline by no more than seven per cent. Given the fact that the present birth rate is about thirty-seven per thousand, this would hardly make a dent in the population problem. However, this is much too pessimistic a conclusion. This becomes clear when we look at the intercept multipliers in Table 2.[18] They measure the effect of the endogenous variables on the birth rate. Thus the intercept multiplier with respect to infant mortality rate is 0.71. This means that a downward shift in the infant mortality intercept of one per cent will cause a 0.71 per cent reduction in the birth rate. Thus, given the fact that the Indian infant mortality rate is about 137.0 per 1000, it would appear that there is considerable scope for reducing birth rate by reducing infant mortality. We may thus conclude from this that the role of reducing infant mortality should be regarded as a part of policies for controlling population growth and hence more attention should be paid to measures leading to reduced infant mortality.

It is interesting to compare some of the elasticity multipliers and intercept multipliers for India and the United States of America -- the only country for which such estimates are available. The relevant information is given in Tables 3 and 4.

Given the somewhat different specifications of the two models and the data used (time series for the U.S. and cross-section for India) I do not want to exaggerate the importance of this comparison. Nevertheless the differences are far too striking to be a mere coincidence. Two differences stand out rather glaringly: (a) a much more pronounced role of education in the United States than in India--all that it probably means is that the quality of education and efficiency in its use leave much to be desired in India; and (b) while there is virtually no impact of changes in income and female participation rate on birth rate in India, just the opposite is the case for the United States. In so far as the effect of infant mortality is concerned, the results for India are just the reverse of the United States. This latter difference -- (b) -- would appear to support Schultz (1973) and Gregory et al.'s(1972 arguments. It also shows that the hypothesis about the opportunity cost of women's time being a determining factor of birth rate, at least in the Indian conditions would appear to be not very strongly applicable. We should, however, emphasize that even for India we cannot dismiss it as being irrelevant because in the structural equation we did get a negative coefficient for the female participation rate, which is at least suggestive. It may well be that a separate treatment of urban birth rate will provide a stronger support for the Mincer-Schultz hypothesis.

IV. CONCLUDING REMARKS

Our simultaneous equations model captures some important aspects of India's fertility behaviour. In terms of the signs of various coefficients, the model performs quite well. In terms of the policy implications we find that literacy has a significant effect though its direct impact is very small. The best chances of reducing the birth rate would appear to be through a reduction of infant mortality where literacy plays a significant role.

Given the limitations of the data, the results of this must of necessity be regarded as tentative. It would be useful to estimate a more disaggregated model, say, for example in terms of rural-urban dichotomy. It would, of course, be highly desirable if a time series study could be carried, but this would appear to be almost impossible at the present time. The only hope of a more thorough study thus is the use of more diversified cross-section data.

APPENDIX

The data, as already pointed out, relate to the year 1961 and cover thirteen states. Thus in all there are thirteen observations. The sources are given below.

1. Data on birth rates, population, infant mortality rates, dependency rate, age at marriage of females, percentage of male labour force in agriculture, were collected from Ashish Bose (ed.), Patterns of Population Change in India 1951-61, Calcutta: Allied Publishers, 1967.
2. Data on female participation rates, literacy rates, population density, and female literacy rates, were collected from V.G. Kulkarni, Statistical Outline of Indian Economy, Bombay: Vora and Company, 1968.
3. Data on per capita income were collected from the National Accounts of Less

Developed Countries, Paris: O.E.C.D., 1968.

4. Data on per capita health expenditure were collected from the Health Statistics of India, 1962, Government of India.

5. Data on per capita energy consumption were collected from the Demand for Energy in India, 1960-1975, National Council of Applied Economic Research, New Delhi. This book provides data by end use and for various sources of energy for each state. Different sources of energy were reduced to KWH using the conversion factors given in this book and then added to give total energy consumption. Using state population as a denominator, we calculated per capita energy consumption for each state.

TABLE 1

ELASTICITY MULTIPLIERS

Endogenous Variables	Exogenous Variables						
	LR	AMF	HE	FLR	LA	ZD	KP
BR	-0.0399	-0.1457	-0.0197	-0.0016	-0.0113	-0.0012	0.0152
IMR	-0.3092	-0.0146	-0.0527	-0.0002	-0.0011	0.0298	-0.3752
Y	0.4245	0.0217	0.0029	0.0002	0.0017	-0.0443	0.5579
FPR	0.1175	0.4292	0.0580	0.5672	4.1550	0.0036	-0.0447
ZT	-0.0226	-0.0824	-0.0112	-0.0009	-0.0064	-0.0007	0.0086

TABLE 2

INTERCEPT MULTIPLIERS

Birth rate with respect to

Y	0.080
FPR	-0.002
IMR	0.710

TABLE 3

ELASTICITY MULTIPLIERS

(Effect of education)

Endogenous variable	India	U.S.
BR	-0.0399	-0.496
Y*	0.4245	1.018
FPR	0.1175	0.260
IMR	-0.3092	-2.180

*For the U.S. this relates to permanent income. The U.S. estimates are from Gregory et al.(1972).

TABLE 4

INTERCEPT MULTIPLIERS

Birth rate with respect to

Endogenous variable	India	U.S.
Y*	0.080	0.792
FPR	-0.002	-0.684
IMR	0.710	-0.086

*See footnote in Table 3. Source as for U.S. as in Table 3.

FOOTNOTES

1. I should like to thank Professor P. Krishnan for his very helpful suggestions. My thanks also go to J. Bosma, Alan Sharpe and Hugh Williams for their assistance with computations.

2. Bhagwati (1966), p. 196.

3. It is of course understood that the choice of endogenous variables to some extent is always arbitrary. Thus one could always argue that a number of other variables, like per capita energy consumption, the percentage of male labour force in agriculture, and the age at marriage should also be considered as endogenous. However, given the paucity of data and the small number of observations at our disposal, the model is not expanded to allow for additional endogenous variables. But, as shown below, the present model, by allowing interdependence between income, female participation rate, birth rate, and infant mortality rate does

throw more light on decisions relating to fertility than single equation models.

4. Schultz (1969), p.155. See also Mincer(1963) and Cain(1966).

5. See Kim(1969) for the use of such measure.

6. More specifically, the argument is as follows. Let $Q = f(K,L)$ where Q is GNP, K is capital and L is labour. Assuming a linear homogeneous production function, we get

   $\frac{Q}{N} = F(\frac{K}{N}, \frac{L}{N})$ where N is population.

   L/N is total labour force participation rate.

   Then $\frac{L}{N} = f(ZT)$ $f' < 0$ and

   $\frac{Q}{N} = F[\frac{K}{N}, f(ZT)] = F_1(\frac{K}{N}, ZT)$ and

   $F_{1ZT} < 0$.

7. See also Enke (1973) on how increasing fertility by changing the dependency rates lowers per capita income. For the use of a somewhat similar equation, see Gregory et al.(1972).

8. See Cain (1966, 1973) and Benham (1971).

9. See United Nations (1962).

10. See Farooq (1972).

11. Leff (1969), p. 887.

12. We recognize the problems of aggregation involved in the transition from the model based on the household to the states, see Gupta (1969). But to repeat these problems here would amount to no more than a ritualistic exercise, for given the limitations of the data, there is little that we can do to remedy these problems.

13. Needless to say that there are obvious objections to the use of such small sizes. But since in this case the only other alternative was that I abandon this study, I decided to proceed inspite of the limitations inherent in small samples. It should, however, be emphasized that for the less developed countries, data limitations is a chronic problem and therefore we must make do with whatever information we have.

14. Schultz (1973). He reports such an association for Bangladesh, Puerto Rico, Taiwan, Chile, and the Phillipines.

15. Schultz (1973), p. 73.

16. See Gregory et al. (1973).

17. Gupta (1971).

18. The methodology of the intercept multipliers is due to Gregory et al. (1972).

REFERENCES

1. Irma Adelman, "An Econometric Analysis of Population Growth", American Economic Review, 53 (June 1963), 314-39.

2. Gary S. Becker, "An Economic Analysis of Fertility", in Demographic and Economic Change in Developed Countries. Universities-National Bureau Conference Series 11. Princeton, N.J.: Princeton University Press, 1960.

3. Lee Benham, "The Labor Market for Registered Nurses: A Three Equation Model", Review of Economics and Statistics, 53, (1971), 246-52.

4. Jagdish Bhagwati, The Economics of Underdeveloped Countries, New York: McGraw-Hill Book Company, 1966.

5. Glen G. Cain, Married Women in the Labor Force: An Economic Analysis, Chicago: University of Chicago Press, 1966.

6. _____ and Adriana Weininger, "Economic Determinants of Fertility: Results from Cross-Sectional Aggregate Data", Demography, 10, (May 1973), 205-24.

7. Edwin D. Driver, Differential Fertility in Central India, Princeton: Princeton University Press, 1963.

8. Stephen Enke, "Population Growth and Economic Growth", The Public Interest, No. 32, (Summer 1973), 86-96.

9. G. M. Farooq, "An Aggregative Model of Labor Force Participation in Pakistan", The Developing Economies, (September 1972), 267-89.

10. Paul R. Gregory, John M. Campbell and Benjamin Cheng, "A Cost-Inclusive Simultaneous Equation Model of Birth Rates", Econometrica (40), No. 4., (July 1972), 681-88.

11. _____, "A Simultaneous Equation Model of Birth Rates in the United States", The Review of Economics and Statistics, 54, (Nov. 1972).

12. _____, "Differences in Fertility Determinants: Developed and Developing Countries", Journal of Development Studies, 9,(January 1973), 233-42.

13. K. L. Gupta, "Dependency Rates and Savings Rates: Comment", American Economic Review, 61, (June 1971), 469-71.

14. _____, Aggregation in Economics,

Rotterdam, 1969.

15. Y. C. Kim, "Sectoral Output-Capital Ratios and Levels of Economic Development: A Cross-Section Comparison of the Manufacturing Industry", Review of Economics and Statistics, 51(1969) 453-58.

16. N. H. Leff, "Dependency Rates and Savings Rates", American Economic Review, 59, (December 1969), 886-96.

17. Jacob Mincer, "Market Prices, Opportunity Costs and Income Effects", in Measurement in Economics, Stanford: Stanford University Press, 1963.

18. H. B. Okun, "The Firth Rate and Economic Development: A Comment", Econometrica, 33, (1965), 245.

19. Paul T. Schultz, "An Economic Model of Family Planning and Fertility", Journal of Political Economy, 77, No. 2, (March/April 1969), 153-80.

20. _____ , "A Preliminary Survey of Economic Analyses of Fertility", American Economic Review, 63, (May 1973), 71-78.

21. United Nations, Department of Economics and Social Affairs, Demographic Aspects of Manpower, Report I: Sex and Age Patterns of Participation in Economic Activities, Population Studies, No. 33, New York, 1962.

22. Rober J. Willis, "A New Approach to the Economic Theory of Fertility Behaviour", Journal of Political Economy, 81, No. 2, Part II, (March/April 1973), 514-64.

DEVELOPMENT OF A SCALE DESIGNED TO MEASURE FUNCTIONAL DISTANCE VISION LOSS USING AN INTERVIEW TECHNIQUE

Kenneth W. Haase and E. Earl Bryant
National Center for Health Statistics

Introduction

The National Health Interview Survey, in addition to providing health statistics on the population of the United States, carries out a research program designed to improve or to develop new survey methodologies. This paper presents the findings of one of the recent survey research activities conducted by the National Center for Health Statistics in cooperation with the American Foundation for the Blind and the National Society for the Prevention of Blindness. The purpose of this study was to develop and test three scales designed to measure functional vision loss by use of an interview technique. The scales consisted of a distance vision scale, a near vision scale, and a self-evaluation scale related to trouble seeing. This paper presents a preliminary assessment of the distance vision scale when used alone and when used in conjunction with the self-evaluation scale.

Methodology and Study Design

The basic methodology for the study involved the collection of data from two sources - an interview with clinic patients and an eye examination by ophthalmologists and clinic technicians which was performed immediately following the interview.

The universe consisted of patients 6 years of age and over who visited the general receiving wards of six eye clinics[1] during a four to six week period beginning in December 1972. Patients visiting the clinics for the first time were excluded.

Differential sampling rates were applied by strata and clinic such that the expected total sample size would be about the same for each clinic and for each of four visual acuity classes (better than 20/50, 20/50 to better than 20/100, 20/100 to better than 20/200, and 20/200 or worse). The sample consisted of 1,726 patients of whom 1,661 responded in the study.

Characteristics of the Sample Population

A most important qualification of the data presented in this paper is that they are applicable to a very select population, one which contains a large proportion of visually impaired, elderly, and poorly educated people. Numerous studies have indicated that the elderly and the less educated often have problems responding in interview surveys. Therefore, these factors should be considered when interpreting the findings of this study.

Development and Analysis of the Distance Scale

In development of the distance scale, major consideration was given to the types of questions that would identify persons with functional distance vision loss and could discriminate between various degrees of that loss.

The scale consisted of the five questions shown in Figure A. The questions are ordered in the form of a Guttman Scale[2]; that is, the first four questions are ordered so that when the first negative answer is obtained, all following answers are expected to be negative. The Guttman technique permits the use of several approaches in evaluating the merits of this instrument. These include the assessment of face validity, construct validity and content validity.

Figure A. The distance scale questions used in the vision study.

(1) (When wearing glasses) can you <u>see</u> well enough to recognize a friend if you get close to his face?

(2) (When wearing glasses) can you <u>see</u> well enough to recognize a friend who is an arm's length away?

(3) (When wearing glasses) can you <u>see</u> well enough to recognize a friend across a room?

(4) (When wearing glasses) can you <u>see</u> well enough to recognize a friend across a street?

(5) Do you have any problems seeing distant objects?

Face validity, while somewhat subjective, should be the first criterion applied to any survey technique. The questions applied to this scale were "Do these questions make sense in classifying functional vision loss?" and "Do they form a hierarchy of severity?" Since the reference point, "recognizing a friend", was kept constant and the conceptual stimulus was decreased by moving the friend further from the respondent, the scale has the appearance of logically classifying various degrees of functional vision loss.

In terms of construct validity the Guttman approach permits a measurement of internal consistency within the scale itself. Each sample person was asked all of the first four questions of the scale regardless of the previous answer.

For example, if a person reported he could not see a friend across a room he was still asked whether he could see a person across a street. Therefore, it was possible to determine scaleability by analyzing the consistency of the responses. Of the 1,661 persons who answered these questions only about 1 percent responded inconsistently. Based on experience from other studies involving scaleability these findings indicate a very high degree of consistency.

The final measurement of validity is content validity; that is, whether the scale actually measures what it is intended to measure. However, before looking at the findings which compare interview data with clinical measurements, we should give some attention to the differences between these two measuring techniques. How a person perceives he can function is related to a number of factors of which his physical capability is only part. These scales are psychological measurements which will be influenced by actual visual acuity measurements. Also they will be related to the patient's own subjective evaluation of the severity of his visual impairment and the degree of effort he puts forth in overcoming it. In addition the environment in which the person generally functions may be quite different from the clinic environment in which the examination was performed. Therefore, both measurements, assuming that they adequately represent the phenomenon of interest, are important statistics in their own right. Since the two measurements are different we do not expect a perfect association, but since they both measure the same phenomenon from a different perspective, we should expect to find a statistical relationship. In this paper we have used Pearson's phi coefficient[3] as an indicator of the degree of association between the two measurements.

If one accepts the hypothesis that persons with similar visual acuity measures can have different perceptions of their degree of functional vision loss, how then does one interpret a statistical correlation between the two measures. To some degree it must be a value judgement. But, comparisons must also be made among the different subgroups, identified by this scale to determine if the distributions of these subgroups by visual acuity are different and if these differences are in the directions expected. Further, the analysis should also include an analysis of the outliers. While we might accept the fact that a person's perception of his degree of functional vision loss can vary considerably with the measurement of his visual acuity, we could evaluate the scale in terms of the apparent inconsistencies. For example, a person who is classified by the scale as having a severe vision loss, should not be expected to have a normal or near normal visual acuity. These outliers will be referred to in this paper as potential false positives and potential false negatives.

Table 1 presents the distribution of the sample according to visual acuity by degree of functional distance vision loss as measured by the distance scale. The clinical measures in this table as in all the following tables are based on measures of visual acuity in the best eye with the sample persons using the type of corrective lens that he usually uses.

Table 1. Number and Percent Distribution of Sample Persons According to Visual Acuity (present corrections in best eye) by Distance Vision Scale.

VISUAL ACUITY (present correction in best eye)	Total		DISTANCE VISION SCALE				
			Cannot recognize a friend at arm's length away	Can recognize a friend at arm's length but not across a room	Can recognize a friend across a room but not across a street	Can recognize a friend across a street	
	Persons[1]	Percent				Some problem seeing distant objects	No problem seeing distant objects
			(Percent Distribution)				
TOTAL	1576	100.0	12.9	16.0	19.8	14.7	36.6
20/400 or worse	219	100.0	53.0	26.5	11.4	3.7	5.5
20/200 to better than 20/400	178	100.0	15.7	36.5	28.1	9.6	10.1
20/70 to better than 20/200	225	100.0	9.3	23.6	32.0	15.1	20.0
20/40 to better than 20/70	339	100.0	7.9	14.8	25.4	16.8	35.1
20/25 to better than 20/40	312	100.0	2.6	7.1	16.4	19.9	54.2
Better than 20/25	303	100.0	1.0	1.3	9.2	17.8	70.6

[1]Excluded from this table are 28 persons for whom the distance scale measure was unknown and 57 for whom visual acuity was unknown.

The value of phi for the distribution for the sample in this table is .35. This somewhat weak association (the value of phi can range from 0 to 1) is partially due to the difference in the measures as discussed above. Also, it may be influenced by how the data are grouped. Since we have no prior evidence to indicate what scale score should be expected for a given visual acuity group, determining adequate cutting points is somewhat difficult. It can be observed, however, that for the extreme visual acuity groups, i.e. 20/400 or worse and the groups better than 20/40, there is a tendency to cluster around the scale scores that might be expected for these groups. However, for the middle visual acuity groups the distribution shows a much wider variation with no salient modal measure.

In analyzing the outliers, that is, those observations that appears to be inconsistent, we found that for the 20/400 or worse group an accumulative total of 9.2 percent reported that they could see well enough to recognize a friend across a street. Ten percent of persons with 20/200 to better than 20/400 vision and 20 percent of persons with 20/70 to better than 20/200 vision reported having no trouble seeing. For those persons with good or normal vision (better than 20/25), an accumulative total of 11.5 percent reported they could not see a friend across a street.

Table 2 shows how the sample is distributed by visual acuity measurement for each of the scale categories. Of the 203 persons who reported that they could not see well enough to recognize a friend an arm's length away 57 percent had a visual acuity of 20/400 or worse, while 1.5 percent had normal vision (better than 20/25), and an accumulative total of 5.4 percent had a visual acuity of better than 20/40. At the other end of the scale, of the 577 persons who reported that they had no problem seeing distant objects, an accumulative total of 5.2 percent had a visual acuity of 20/200 or worse, which is the cutting point for determining legal blindness.[4] As expected the bulk of those persons reporting no problems are clustered in the better visual acuity groups.

The vision questionnaire included a set of questions designed to obtain the respondent's self-evaluation of his vision in each eye separately (see Figure B).

Figure B. The self-evaluation scale questions used in the vision study.

(1) (When wearing glasses) how much trouble do you have seeing with your left eye--a lot of trouble, a little trouble, or no trouble at all?

(2) (When wearing glasses) how much trouble do you have seeing with your right eye--a lot of trouble, a little trouble, or no trouble at all?

Table 2. Number and Percent Distribution of Sample Persons According to the Distance Vision Scale by Visual Acuity (present corrections in best eye).

VISUAL ACUITY (present correction in best eye)	TOTAL	DISTANCE VISION SCALE				
		Cannot recognize a friend at arm's length away	Can recognize a friend at arm's length but not across a room	Can recognize a friend across a room but not across a street	Can recognize a friend across a street	
					Some problem seeing distant objects	No problem seeing distant objects
Total Number of Persons[1]	1576	203	252	312	232	577
		(Percent Distribution)				
Total Percent	100.0	100.0	100.0	100.0	100.0	100.0
20/400 or worse	13.9	57.1	23.0	8.0	3.5	2.1
20/200 to better than 20/400	11.3	13.8	25.8	16.0	7.3	3.1
20/70 to better than 20/200	14.3	10.3	21.0	23.1	14.7	7.8
20/40 to better than 20/70	21.5	13.3	19.8	27.6	24.6	20.6
20/25 to better than 20/40	19.8	3.9	8.7	16.4	26.7	29.3
Better than 20/25	19.2	1.5	1.6	9.0	23.3	37.1

[1]Excluded from this table are 28 persons for whom the distance scale measured was unknown and 57 patients for whom visual acuity was unknown.

These questions provide a four point scale of the respondents self assessment of his ability to see with each eye ranging from blind to no trouble seeing.

Although the respondents were instructed to respond to the distance scale in relation to their total vision, it is possible to hypothesize that some persons with an impairment in only one eye might respond in terms of that eye rather than their overall vision. In a somewhat similar study designed to develop a hearing scale[5] a relatively large segment of the false positives resulted from persons with little or no hearing loss in one ear who were responding in terms of their impaired ear.

To test whether this phenomenon was also present in the distance vision scale we combined the responses obtained for each person's self-evaluation for each eye to establish two categories: (1) those persons reporting at least a little trouble seeing in both eyes and; (2) those persons reporting they have no trouble seeing in at least one eye. Although there will be some reduction in the field of vision, a person who has severe vision loss in one eye but normal vision in the other should be able to see well enough to recognize a friend. Therefore, persons reporting no trouble seeing in one eye are treated as a separate group and only those persons with some trouble seeing in both eyes are classified according to their response to the distance scale. Table 3 shows how the sample is distributed by visual acuity according to this joint classification.

The phi coefficient for Table 3 is .36 which is similar to the association observed in the first set of tables. However, there is a shift in the potential outliers. Using the distance scale by itself we found that all but 9.2 percent of persons with 20/400 or worse reported that they could not see a friend across a street. With this joint classification 17.6 percent of this severe visual acuity group are potential false negatives, of which the bulk fall into the category of one or both eyes good. A similar increase of potential false negatives is also observed in those groups with 20/70 or worse. It would appear that some proportion of those persons who report that they have no trouble seeing in one eye do in fact have severe vision loss in their better eye. It is possible that because of the subjective nature of the self-evaluation scale, some respondents with extreme loss in one eye and the other eye impaired, but to a lesser degree, may overrate their better

Table 3. Number and Percent Distribution of Sample Persons According to Visual Acuity (present correction) by Self-Evaluation and Distance Scale Measures.

VISUAL ACUITY (present correction in best eye)	SELF-EVALUATION AND DISTANCE VISION SCALE							
	Total		Trouble Seeing in Both Eyes					
			Cannot recognize a friend at arm's length away	Can recognize a friend at arm's length but not across a room	Can recognize a friend across a room but not across a street	Can recognize a friend across a street		No trouble seeing in one or both eyes
						Some problem seeing distant objects	No problem seeing distant objects	
	Persons[1]	Percent	(Percent Distribution)					
TOTAL	1526	100.0	11.7	13.2	13.4	7.7	9.3	44.8
20/400 or worse	215	100.0	50.7	22.8	8.8	2.3	1.9	13.5
20/200 to better than 20/400	168	100.0	15.5	32.7	24.4	7.1	5.4	14.9
20/70 to better than 20/200	213	100.0	9.4	20.7	23.5	9.9	8.0	28.6
20/40 to better than 20/70	324	100.0	5.9	11.7	15.4	10.8	12.4	43.8
20/25 to better than 20/40	306	100.0	1.3	4.6	9.5	11.1	13.4	60.1
Better than 20/25	300	100.0	.3	.3	5.0	3.3	10.3	80.7

[1]Excluded from this table are 135 persons for whom the self-evaluation scale, distance scale or visual acuity measures were unknown.

eye because this judgement is made relative to their worse eye. Although we plan to test this hypothesis in future analysis, at the present time we can only speculate on the reasons for these apparent inconsistencies.

While combining the self-evaluation scale with the distance scale increases the potential false negatives, it does appear to decrease the proportion of potential false positives. Using the distance scale alone we saw that an accumulative total of 11.5 percent of the persons with normal vision reported that they were unable to recognize a friend across a street. By excluding those persons who reported having no trouble seeing in one or both eyes the proportion of potential false positives is reduced by 5.6 percentage points. Therefore, if we assess the distance scale as a screening device, the inclusion of the self rating scale appears to decrease its sensitivity in that it increases the proportion of potential false negatives but increases its specificity in that it decreases the proportion of potential false positives. Since in the general population only a small proportion of persons will have a vision problem, the false positives will cause much more distortion of an estimate derived from these procedures than would be caused by false negatives. In fact, if the 5.6 percent potential false positives are actually false positives, and if the same proportion were present in a national survey, the estimate for vision impairment would be doubled. However, there are reasons to assume that the proportion of false positives within a general population would not be of this magnitude. First, some of these potential false positives may be caused by other vision defects such as restricted field vision which may not be reflected in the visual acuity measurement. Although information on other vision defects is available to us, we have not had time to analyze it. Secondly, since all sample persons when interviewed were visiting a clinic for some reason related to their eyes or vision there might have been a tendency for some proportion of the study population to exaggerate their vision problem.

Table 4 presents the number of persons classified by the joint distance and self-evaluation scale distributed according to their visual acuity measures. By excluding the persons who report no trouble seeing in one or both eyes the proportion of persons with normal acuity is decreased in all of the distance scale groups. The proportion of persons with a visual acuity of 20/70 or worse is increased in all the scale response groups including those who report no problem seeing by this joint classification.

Table 4. Number and Percent Distribution of Sample Persons According to Self-Evaluation and Distance Scale Measures by Visual Acuity (present correction).

VISUAL ACUITY (present correction in best eye)	SELF-EVALUATION AND DISTANCE VISION SCALE						
	Total	Some Trouble Seeing in Both Eyes					No trouble seeing in one or both eyes
		Cannot recognize a friend at arm's length away	Can recognize a friend at arm's length but not across a room	Can recognize a friend across a room but not across a street	Can recognize a friend across a street		
					Some problem seeing distant objects	No problem seeing distant objects	
Total number of persons[1]	1526	179	201	204	117	142	683
Total Percent	100.0	100.0	100.0	100.0	100.0	100.0	100.0
20/400 or worse	14.1	60.9	24.4	9.3	4.3	2.8	4.3
20/200 to better than 20/400	11.0	14.5	27.4	20.1	10.3	6.3	3.7
20/70 to better than 20/200	14.1	11.2	21.9	24.5	18.0	12.0	8.9
20/40 to better than 20/70	21.1	10.6	18.9	24.5	29.9	28.2	20.8
20/25 to better than 20/40	20.1	2.2	7.0	14.2	29.1	28.9	26.9
Better than 20/25	19.6	.6	.5	7.4	8.6	21.8	35.4

[1]Excluded from this table are 135 persons for whom self-evaluation scale, distance scale or visual acuity measures were unknown.

In summary, the analysis of responses to the distance scale indicates a high degree of internal consistency, which provides evidence that the order and nature of this set of questions have ordinal characteristics.

When comparing the responses from the distance scale with visual acuity measures, we found a positive but relatively weak statistical association. While combining the distance and self-evaluation scale increased the proportion of potential false negatives, it decreased the proportion of potential false positives, which is assumed to create a more important measurement problem. Although there remains a number of unexplained inconsistencies in these findings, some of which might be explained in further analysis of these data, we are generally encouraged by the distance scale's ability to classify populations according to perceived functional vision loss. Therefore, we are presently planning to incorporate this scale into the next cycle of the National Health Examination Survey to test it on the national population. The methodology will be similar to that employed in this study but the findings can be inferred to the general population.

ACKNOWLEDGEMENTS

The authors wish to thank the following persons for their help in the planning and conduct of the Vision Study: Dr. Milton Graham and Robert Robinson, American Foundation for the Blind; Elizabeth Hatfield, National Society for the Prevention of Blindness; Dr. Arthur Keeney, University of Louisville; Dr. Alfred Sommer, Wilmer Institute; Dr. Donald Doughman, University of Minnesota; Dr. Morton Smith, Washington University; Dr. Joel Kraut and E. J. Stockman, Jr., Massachusetts Eye and Ear Infirmary; James McGee, Wills Eye Hospital and Research Institute; E. A. Petrelli, New York Eye and Ear Infirmary; and Abigail Moss, Robert Wright, Beany Slater and Nelma Keen, National Center for Health Statistics.

REFERENCES

[1]The clinics participating in this study were (1) Massachusetts Eye Infirmary, Boston, Mass.; (2) New York Eye and Ear Infirmary, New York, N.Y.; (3) University of Minnesota Eye Clinic, Minneapolis, Minn.; (4) Washington University Medical Center, St. Louis, Mos.; (5) Wills Eye Hospital, Philadelphia, Pa.; and (6) Wilmer Clinic, Baltimore, Md.

[2]Guttman, L., "The Cornell Technique for Scale and Intensity Analysis," Educational and Psychological Measurements, 1957, 7:247-279.

[3]Kendall, M. G. and Stuart, A., The Advanced Theory of Statistics, Vol. 2, Hafner Publishing Co., 1961.

[4]In clinical terms, legal blindness is usually defined as central visual acuity of 20/200 or less in the better eye, with correcting glasses; or central visual acuity of more than 20/200 if there is a defect in which the peripheral field has contracted to such an extent that the widest diameter of vision subtends an angular distance no greater than 20 degrees.

[5]National Center for Health Statistics. Vital and Health Statistics, Series 2, No. 37, "Development and Evaluation of an Expanded Hearing Loss Scale Questionnaire," PHS Pub. 1000, Public Health Service, Washington, U.S. Government Printing Office, April 1970.

CONTRACEPTION STRATEGIES AND FERTILITY OUTCOMES: ESTIMATES OF A STOCHASTIC MODEL OF REPRODUCTION

James J. Heckman, The University of Chicago and
The National Bureau of Economic Research

Robert J. Willis, City University of New York and
The National Bureau of Economic Research*

In this paper we report some results from an ongoing study in the economics of fertility. Space limitations preclude a complete summary of our work. Accordingly, in this paper we briefly sketch the economic model and discuss how it differs from previous work in economic demography. We then discuss a statistical problem that arises in estimating the model and illustrate the practical importance of this issue using data from the 1965 Princeton National Fertility Study. For a more complete description of our work, readers are referred to Heckman and Willis, 1974.

I
The Model

Previous work in economic demography assumes that families choose a desired number of children and a desired amount of child quality in a world of perfect foresight. This approach, developed most fully in Willis (1973), neglects the uncertainty inherent in the fertility process. In this paper a family is viewed as controlling parameters of a stochastic birth process through choice of contraceptive techniques and levels of use effectiveness. There are monetary, time, and psychic costs associated with each form of contraception and level of use effectiveness. Families are assumed to maximize expected utility over a finite horizon, and contraception decisions are made with this view in mind.

A family's decision problem may formally be represented as a dynamic programming problem. Except in simple cases, a complete analytical solution to the problem is unavailable. Nonetheless certain insights do emerge from the analysis: (1) the fertility strategy of a family, and its outcome, depend on its previous history of fertility outcomes, including realized spacing intervals. (2) The stochastic models of reproduction advanced by Perrin and Sheps (1964) are embedded in a choice theoretic framework. (3) Given sufficient data on the time series of a family's income and wage rates, testable hypotheses may be generated about the life cycle history of contraceptive choice and fertility outcomes. (4) Timing, spacing, and final number of children are all generated by a common probability process that can be altered by contraceptive decisions.

II
A Statistical Problem

The model of the previous section immediately suggests that the family's fertility decisions may be represented by a decision tree. Conditional on a sequence of realized events, families make fertility decisions. The probabilities in the observed stochastic birth process may be parameterized, and hypotheses about the effect of such economic variables as the education of the wife, her age at marriage, and the husband's income may be tested by determining at what stages, and in which decisions, economic variables can be said to contribute anything to explaining fertility outcomes. Even in the absence of a fully developed theory of family planning, estimates of the constituent probabilities allow us to account for the importance of economic variables in the various components of the birth process.

In estimating these probabilities from a random sample of individuals, it is important to note that unless very strong statistical assumptions are made, the simple semi-Markov probability structure does not lead to a simple likelihood function in which estimated parameterized probabilities can accurately be said to predict the probabilities of observed events for individuals. To see that this is so, it is important to distinguish three sources of variation in observed birth intervals among individuals: (1) purely random factors that arise independently in each time period, and are independent of random factors in other time periods, (2) random factors, including unobservable variables, that are correlated across time periods, (3) deterministic variables such as income and education that can be measured, and which are assumed to affect the probabilities.

To fix ideas, suppose we are concerned solely with estimating the parameters of the probability process determining whether a woman has a first pregnancy. Inherent in the model is the notion of a time series of events. A woman has a first pregnancy in month j only if she has not had a first pregnancy in months $1,\ldots,j-1$. The most general way to model this probability is to imagine a set of continuous random variables S_1, $S_2,\ldots$, which may be thought of as index functions. The S_i, $i=1,\ldots,\infty$, are assumed to be intercorrelated. The event of a woman becoming pregnant in the first interval depends on what value the "wheel of chance" throws up for S_i. Suppose that her education E is the only economic variable of interest. We may then define $\alpha_o+\alpha_1 E$ so that if $S_1<\alpha_o+\alpha_1 E$ a woman becomes pregnant in the first interval and leaves the sample while if the inequality is reversed, the woman is not pregnant and stays in the sample. The probability of a woman becoming pregnant in the jth interval is thus

$$(1)\quad \Pr(S_1>\alpha_o+\alpha_1 E,\ldots,S_{j-1}>\alpha_o+\alpha_1 E_1,S_j<\alpha_o+\alpha_1 E)\ .$$

If we assume that the S_i are independently and identically distributed, this probability may be written as

$$(2)\quad \prod_{i=1}^{j-1}\Pr(S_i>\alpha_o+\alpha_1 E)\ \Pr(S_j<\alpha_o+\alpha_1 E)\ .$$

If each S_i is assumed to be distributed normally with mean zero, and variance σ_s^2, the probability statement may be written using probit functions as

$$(3)\quad \left[\int_{\frac{\alpha_o+\alpha_1 E}{\sigma_s}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}dt\right]^{j-1} \int_{-\infty}^{\frac{\alpha_o+\alpha_1 E}{\sigma_s}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}dt$$

If the S_i were assumed to be logistically distributed, a similar probability statement using cumulative logistics could easily be written.

If the S_i for all women are generated by the same random process, we may use the principle of maximum likelihood to estimate $\frac{\alpha_o}{\sigma_s}$ and $\frac{\alpha_1}{\sigma_s}$ by taking a sample of women with different birth intervals, and choosing parameter values which maximize the probability of observing the sample distribution of birth intervals.

Note, however, a crucial step in the argument. We assumed that over time, the S_i were independently distributed. This assumption rules out serial correlation in the S sequence. Such serial correlation may naturally arise if there are unmeasured random variables which remain at, or near the same level, over time for a given individual, but which are randomly distributed among individuals. For example, unmeasured components of fecundability (e.g. semen counts of husbands, tastes for coital activity, and variations in contraceptive efficiency) plausibly have a persistent component for the same individual across time periods although these components may vary widely among individuals.[1] Similarly, important economic variables may be missing in a given body of data.[2]

Following a convention in the analysis of covariance, we may decompose S_i into two components

$$(4)\quad S_i = U_i + \varepsilon$$

where U_i is a random variable with mean zero and variance σ_u^2, and ε is a random variable with mean zero, and variance σ_ε^2. We further assume, letting "E" be the mathematical expectation, that

$$(5)\quad E(U_i U_j) = 0, \quad i \neq j$$

$$E(U_i \varepsilon) = 0 \quad .$$

Then S_i is a random variable with mean

$$E(S_i) = 0$$

and

$$E(S_i S_j) = \sigma_\varepsilon^2 \quad i \neq j$$

$$(6)\qquad = \sigma_\varepsilon^2 + \sigma_j^2, \quad i=j \quad ,$$

so that the correlation coefficient between S_i in any two periods, ρ, may be defined as

$$(7)\quad \rho = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_u^2} \quad .$$

Clearly, it is possible to imagine more general intercorrelation relationships such as a first-order Markov process. These generalizations are straightforward and, since they are not of direct interest in this paper, are not pursued here. If intercorrelation applies because there are persistent omitted variables, the probability of a woman becoming pregnant in interval j can no longer be written in the simple form of equation (2) (or if S is assumed normal, as in equation (3)). To see what the appropriate probability statement becomes, note that in general we may write the probability of the event conditional on a given value of ε as

$$(8)\quad Pr(S_1 > \alpha_o+\alpha_1 E, \ldots, S_{j-1} > \alpha_o+\alpha_1 E, S_j < \alpha_o+\alpha_1 E / \varepsilon) \quad .$$

But note that if ε is held fixed, the distribution of S_i conditional on $\varepsilon=\tilde{\varepsilon}$ must satisfy the following properties

$$E(S_i/\tilde{\varepsilon}) = \tilde{\varepsilon}$$

$$(9)\quad E(S_i S_j/\tilde{\varepsilon}) = \tilde{\varepsilon}^2 \quad i \neq j$$

$$= \sigma_u^2 + \tilde{\varepsilon}^2$$

and since the U_i are independent the conditional values of S_i are also independent.

Then we see that

$$Pr(S_1 > \alpha_o+\alpha_1 E, \ldots, S_{i-1} > \alpha_o+\alpha_1 E, S_j < \alpha_o+\alpha_1 E/\tilde{\varepsilon})$$

$$(10) = Pr(S_1 > \alpha_o+\alpha_1 E/\tilde{\varepsilon}) Pr(S_2 > \alpha_o+\alpha_1 E/\tilde{\varepsilon}) \ldots$$

$$Pr(S_j < \alpha_o+\alpha_1 E/\tilde{\varepsilon})$$

so that conditional on $\varepsilon = \tilde{\varepsilon}$ we reach precisely the same functional form as in equation (2) where persistent omitted variables are ignored. However, to solve back to the probability of interest, where ε is permitted to vary between plus and minus infinity, we note that the <u>unconditional</u> probability may be written as

$$\int_{-\infty}^{\infty} Pr(S_1 > \alpha_o+\alpha_1 E/\varepsilon) Pr(S_2 > \alpha_o+\alpha_1 E/\varepsilon) \ldots .$$

$$(11)\qquad Pr(S_j < \alpha_o+\alpha_1 E/\varepsilon) h(\varepsilon) d\varepsilon$$

where $h(\varepsilon)$ is the marginal density function of ε, and ε is permitted to vary over all possible values, as before.

In the special case with S normally distributed with zero mean and variance $\sigma_\varepsilon^2+\sigma_u^2$, equation (11) is seen to be the integral of a multivariate, normal density with equicorrelated variates, and common correlation coefficient ρ. The expression adopted for the computations is

$$(12)\quad \int_{-\infty}^{\infty} \left[\int_{\frac{\alpha^*+\alpha_1^* E+\rho^{1/2} q}{(1-\rho)^{1/2}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}dt\right]^{j-1} \left[\int_{-\infty}^{\frac{\alpha_o^*+\alpha_1^* E+\rho^{1/2} q}{(1-\rho)^{1/2}}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}dt\right] \frac{1}{\sqrt{2\pi}} e^{-q^2/2}dq$$

where $\alpha_o^* = \frac{\alpha_o}{(\sigma_u^2+\sigma_\varepsilon^2)^{1/2}}$ and $\alpha_1^* = \frac{\alpha_1}{(\sigma_u^2+\sigma_\varepsilon^2)^{1/2}}$.

If no serial correlation is present ($\rho=0$), this expression collapses to equation (3). In the more general case, ρ allows us to measure the proportion of total variance in the index S explained by systematic correlated components.

Notice that there is an alternative "incidental parameters" argument that leads directly to equations (11) and (12). Suppose it is argued that in an ordinary probit model a disturbance "ε" appears. This may be viewed as an incidental parameter with density function $h(\varepsilon)$. Following a suggestion of Kiefer and Wolfowitz (1956), the problem of incidental parameters has precisely the general solution written in equation (11) or the specific solution for the normal case as in equation (12).

Yet another interpretation of these results is possible. An individual may be modeled as having a geometric probability process characterizing the probabilities of pregnancy at each interval for a given value of ε. "ε" is, in fact, a random variable governed by a density function $h(\varepsilon)$. Then the true probability of pregnancy at month j is a continuous mixture of geometric processes and is given by equation (11).

Elsewhere we demonstrate that estimates of the population probabilities that neglect the serial correlation phenomenon impose the constraint that the conditional probability of pregnancy in a given month for a group of women with identical economic characteristics is the same for all months from the onset of marriage. If persistent omitted variables are present (i.e. ρ in equation (12) is nonzero), it can be shown that this conditional probability declines with the length of the interval since marriage because more fecund women tend to become pregnant and hence drop out of the sample eligible for a first birth. If ρ is inappropriately assumed to be zero, estimates of the effect of economic variables on the probability of conception are biased.

III
Empirical Results

This section presents estimates of the monthly probability of conception in the first pregnancy interval following marriage using the econometric model discussed in the preceding section. The data consist of a sample of white non-Catholic women, married once with husband present for 15-19 years from the 1965 Princeton National Fertility Study.[3] The sample of all such women was reduced by eliminating women who reported premarital conceptions or who had missing values for relevant variables. The sample was then divided into two groups, contraceptors and noncontraceptors, on the basis of the woman's response to a question concerning the contraceptive methods she used before her first pregnancy (or in her current interval if she had not had a pregnancy). Three variables (wife's education (W), wife's age (A), and husband's predicted income at age 40 (H)) are expected to influence the monthly probability of conception.

Women, in each subsample were "followed" for a maximum of 120 months beginning with their first month of marriage. Among the noncontraceptors we estimate the monthly probability of conception in the first pregnancy interval by maximizing the appropriate likelihood function.[4] That is, using the functional form of the likelihood function implied by (12) we estimate parameters which maximize the likelihood of observing the events that occurred in this subsample. These events are (1) that a given woman conceived in month j (j=1, ..., 120) or (2) that she went 120 months without conceiving. Among the contraceptors, we estimate in similar fashion the monthly probability of conception given that the woman is contracepting. In this case, the events we observe are (1) that a woman conceives in month j while using a contraceptive; (2) that the woman uses a contraceptive for k months without conceiving, at which time she discontinues contraception (this decision is treated as an exogenous event); or (3) she continues using contraception for 120 months and does not conceive.[5]

Parameter estimates for the noncontraceptors are presented in Table 1A and for contraceptors in Table 1B. In each group, we estimated six models which differ in the number of parameters estimated in order to determine the statistical significance of individual parameters or sets of parameters using likelihood ratio tests.

Among these parameters, we have a particular interest in the magnitude of the serial correlation coefficient, ρ, its statistical significance and the influence of its inclusion or exclusion from the econometric model on the other parameters of the model (i.e., the constant term, α_0, and the coefficients of A, W, and H which are, respectively, α_1, α_2 and α_3). Accordingly, we present two estimates of each set of α's in Table 1, one in which ρ is constrained to be zero and one in which ρ is free to assume a nonzero value. The sign of a coefficient indicates the direction of effect of the variable on the probability of <u>not</u> conceiving in a given month. For example, in Table 1 the positive coefficient for α_1 suggests that the later the wife's age at marriage, the less likely she is to conceive in a given month.

In Table 1 ρ is positive and statistically significant in every instance. Among the noncontraceptors $\rho=0.450$ when only the constant term is entered and falls to 0.426 when the wife's age at marriage is held constant, but does not fall any further when wife's education and husband's predicted income are added to the model. Similarly, the estimate of ρ in the contracepting

subsample falls from 0.549 to 0.531 when A is held constant and to 0.526 when W and H are also held constant. If we recall that the definition of ρ is the fraction of persistent variance (σ_ε^2) in total variance ($\sigma_\varepsilon^2+\sigma_u^2$), the decrease in ρ is easily understood as showing that the exogenous variable A in the noncontracepting subsample and the variables A, H and W in the contracepting subsample contribute to the persistent component of variation in conception probabilities among women in the two subsamples. The small size of the decrease in ρ, however, also shows that the contribution of other factors we have not held constant constitutes the major fraction of persistent variation. This suggests that it is unlikely that the heterogeneity problem can be overcome simply by holding constant a number of observable variables.

The size of the decrease in ρ caused by the addition of exogenous variables is, of course, related to the statistical significance of these variables. The wife's age at marriage is the only variable to pass a test of statistical significance at conventional levels in either subsample.

Estimates of the monthly probability of conception and the effects of changes in exogenous variables on that probability differ substantially depending on whether or not serial correlation is taken into account. In Table 2, we present examples of estimates of levels in the monthly probability of conception among noncontraceptors and contraceptors with and without ρ constrained to equal zero. To make the contrasts, we present estimates of the probabilities of conception in the first month after marriage. These estimates are derived from Table 1.

For example, line (1') in Table 1 maps into the second column of line A.1 in Table 2. Similarly, line (1) in Table 1 maps into the first column of line A.1. It is easy to see that the bias from not allowing for serial correlation is quite large. For contraceptors (line A.2 in Table 2) a similar result holds. In both cases, the monthly probability of conception is seriously understated when ρ is constrained to be zero.

In Table 2B, we evaluate the monthly probability of conception for one value of wife's age at marriage for noncontraceptors and contraceptors with and without ρ constrained to be zero from parameter estimates in lines (2), (2'), (5) and (5') in Table 1. Here, we notice that the bias arising from estimates neglecting serial correlation is large.

Similarly, dramatically different estimates for the effect of economic variables on the probability of conception result when allowance is made for serial correlation.

TABLE 1

ESTIMATES OF PARAMETERS OF MODEL FOR CONTRACEPTORS AND NONCONTRACEPTORS IN FIRST PREGNANCY INTERVAL AFTER MARRIAGE

	Constant (α_o)	ρ	Wife's Age at Marriage (α_1)	Wife's Education (α_2)	Husband's* Predicted Income (α_3)	Log e Likelihood
A. Noncontraceptors (177 Observations)						
(1)	2.016					-692.71
(1')	1.214	0.450				-619.50
(2)	1.154		0.0033			-680.42
(2')	0.172	0.426	0.0042			-613.36
(3)	1.022		0.0031	0.017	-0.0033	-679.80
(3')	0.132	0.426	0.0041	-0.004	0.0125	-613.33
B. Contraceptors (246 Observations)						
(4)	2.264					-336.92
(4')	1.780	0.549				-319.43
(5)	1.307		0.0038			-332.42
(5')	0.646	0.531	0.0046			-316.32
(6)	1.072		0.0036	-0.0016	0.0387	-331.82
(6')	0.943	0.526	0.0042	-0.0068	0.0903	-314.89

*This variable is estimated from a regression of husband's income on his education and age, and arbitrarily assigning the value of 40 for age so that the regression prediction is an estimate of husband's permanent income.

TABLE 2

ESTIMATES OF MONTHLY PROBABILITY OF CONCEPTION DERIVED FROM PARAMETER ESTIMATES IN TABLE 1

	(a) Serial Correlation Ignored ($\rho=0$)	(b) Serial Correlation Allowed ($\rho>0$)
A. Model with Constant Term Only (α_0)		
1. Noncontraceptors	.022	.113
2. Contraceptors	.012	.038
B. Effect of Wife's Age at Marriage (Model with α_0, α_1)		
Noncontraceptors		
3. Age 20	.026	.122
Contraceptors		
4. Age 20	.013	.040

Footnotes

*This research was sponsored by a National Institute of Child Health and Human Development grant to the National Bureau of Economic Research. The authors wish to thank A. Dagli and R. Shnelvar for competent programming assistance. This paper has not undergone Bureau staff review and accordingly is not an official publication of the National Bureau.

[1]The problem of heterogeneity is considered in a demographic context by Sheps (1964), Potter and Parker (1964) and Sheps and Menken (1972).

[2]In this paper, we abstract from the further problem that the unobserved components may be correlated with the included variables.

[3]The 1965 National Fertility Study, conducted by Norman B. Ryder and Charles F. Westoff, is a cross-section national probability sample of 5,617 U.S. married women which is described in detail in Ryder and Westoff (1971). For our purposes, its most important characteristics are that it records (retrospectively) the date of marriage of the woman, the dates of each pregnancy termination, the use of contraception in each pregnancy interval, and the time of discontinuation of contraception prior to pregnancy in addition to a number of household characteristics such as income and education.

[4]The methods used are described in Goldfeld and Quandt (1972, Ch. 1). Two algorithms, Powell and GRADX, were used in tandem to ensure that the estimates are stable. That is, in the first stage the parameters of the likelihood function were estimated by the Powell method. These parameters were then given as initial values in a GRADX optimization procedure whose final parameter values are reported in this paper. The computer program, written by C. Ates Dagli and Ralph Shnelvar is available from the authors on request.

[5]Our data only record whether a woman contracepted in a given pregnancy interval and when and if she discontinued contraception. They do not record when she began contracepting or any other interruptions in contraceptions other than the final decision to discontinue.

References

Goldfeld, S.M. and R.E. Quandt, Nonlinear Methods in Econometrics, Amsterdam-London, North-Holland, 1972.

Heckman, J. and R. Willis, "Estimates of a Stochastic Model of Reproduction: An Econometric Approach," in Household Production and Consumption, ed. by N. Terleckyj, Studies in Income and Wealth, Vol. 139, 1974.

Kendall, Maurice and Stuart, Alan, The Advanced Theory of Statistics, Vol. I, New York, Hafner Publishing, 1969.

Kiefer, J. and Wolfowitz, J. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," Annals of Mathematical Statistics, Vol. 27, No. 4, December, 1956.

Pearson, J., "Contributions to the Mathematical Theory of Evolution. Disection of Frequency Curves," Philosophical Transactions of the Royal Society, Series A, Vol. 185, 1894.

Perrin, Edward and Mindell C. Sheps, "Human Reproduction: A Stochastic Process," Biometrics, Vol. 20, March, 1964, pp. 28-45.

Quandt, R., "New Methods for Estimating Switching Regressions," Journal of the American Statistical Association, June, 1972.

Sheps, M.C., "On the Time Required for Contraception," Population Studies 18, 1964, pp. 85-97.

Sheps, M.C. and J. Menken, "On Estimating the Risk of Conception from Censored Data," T.N.E. Greville, Population Dynamics, Academic Press, New York, 1972, pp. 167-200.

Willis, Robert J., "A New Approach to the Economic Theory of Fertility Behavior," Journal of Political Economy, Vol. 81, No. 2, Part II, March/April, 1973, pp. s14-s64.

Zellner, A., "Bayesian and Non-Bayesian Analysis of the Regression Model with Multivariate Student-t Error Terms, May, 1973 (unpublished paper).

Ryder, Norman B., "Contraceptive Failure in the United States," Family Planning Perspectives, Vol. 5, No. 3, Summer 1973, pp. 133-44.

Ryder, Norman B. and Charles F. Westoff, Reproduction in the United States 1965, Princeton: Princeton University Press, 1971.

AVAILABILITY OF FAMILY PLANNING SERVICES IN COLLEGES AND UNIVERSITIES

Gloria Hollis and Karen Lashman
DHEW, HRA, National Center for Health Statistics

I. Introduction

The sexual revolution on the American campus has, in one respect, caused the present generation of college students to be unique. Although, as various studies have pointed out, there has been little change in the sexual behavior of college students since the early twentieth century, there has been a revolutionary extension of sexual freedom long afforded to men to include women as well.

Growing recognition of this fact has led to increased pressure for the expansion of family planning services as an integral part of student health programs in institutions of higher education. For example, the Council on Population of the American Public Health Association (APHA) has proposed that the APHA recommend that college health services offer confidential medical consultation and service on birth control methods, on the diagnosis of pregnancy, and on the diagnosis and treatment of venereal disease.

The American College Health Association (ACHA) has conducted two surveys--one in 1966 and one in 1970--of the practices and policies of college health services in dispensing contraceptives. While the 1966 survey was limited and included only the 458 institutions which were members of ACHA, the 1970 followup survey of 2,558 colleges represented virtually the entire universe of institutions of higher education, including both ACHA (555) and non-ACHA (2,003) members. When the results of these two surveys are compared, it appears that within this short period of time, there has been a substantial increase in the number of colleges providing this service. For example, between 1966 and 1970 the number of college pharmacies prescribing oral contraceptives for unmarried students who had attained majority increased from 4 percent of ACHA responding members or 13 institutions to 59 percent or 158 institutions. This includes 118 ACHA members and 40 non-ACHA member institutions. While in 1966 only 12 colleges reported that they would prescribe contraceptives for unmarried minors, 125 or 47 percent of the responding institutions in the 1970 survey reported the provision of such service. An increase in the number of institutions reporting the provision of contraceptive prescriptions to married students also occurred, with 45 percent of the schools reporting this service in 1966 in contrast to 82 percent of the respondents in 1970. The number of schools which referred students to off-campus physicians for family planning services also seemed to increase significantly--from 6 percent of the respondents in 1966 to 74 percent in 1970.

Although caution must be used in drawing conclusions from these surveys due to their response rates (74 percent in 1966 and 31 percent in 1970), these data nevertheless reflect growing recognition of the need for contraceptive information and services as an integral part of college health programs. They seem to indicate that the most significant change since 1966 was the expansion of services available to unmarried students whether or not they had attained majority. Even if these conclusions are warranted, it is evident that a large number of sexually active unmarried college students--both minors and those who have attained majority--remained unserved in this critical area of need as the 1970's began.

II. Methodology

As part of its responsibility to provide baseline statistics on the health resources of the nation, the National Center for Health Statistics (NCHS) is in the process of compiling a comprehensive list of all facilities in the United States which provide family planning services. This list will provide the basis for a National Inventory of Family Planning Clinics, and the data it contains will be updated through annual surveys.

Many sources were contacted in the compiling of this list. One obvious source was all the colleges and universities in the nation. Therefore, in the spring of 1973, NCHS conducted a mail survey of all college and university health units or infirmaries to ascertain the availability of family planning services on campus.

The universe for this survey consisted of all institutions of higher learning which met the academic and administrative requirements for inclusion in the 1972-73 Directory of the United States Office of Education. The one adjustment made to the universe concerned the 58 Roman Catholic seminaries, which were excluded from the mailout, thereby reducing the total institutions surveyed from 3,042 to 2,984.

The survey utilized a one page letter initially followed by one mail followup to those institutions which did not respond after three weeks. The definition of family planning services for the purposes of this survey was as follows: those medical, social and/or educational services which are primarily concerned with the regulation of conception. Institutions were requested to check the one box in the letter which was most appropriate to them. There were three choices, as follows: (1) no family planning services provided at this institution; (2) students referred to other source for these services; (3) some type of family planning services provided at this institution. Those institutions which did refer students to other sources for family planning services were asked to list these sources as a check on the completeness of the National Inventory of Family Planning Clinics.

The returned letters were sorted into two major groups--providers (those offering family planning services) and nonproviders (those not offering such services). All providers were added to the computer file of family planning facilities being developed by NCHS.

Table 1. Number and percent of colleges and universities providing family planning services, by sex of student body and type of institution.

Type of college or university	Sex of Student Body									
	Total		Male		Female		Coed		Coordinate [1]	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total colleges	2,984	100.0	92	3.1	159	5.3	2,719	91.1	14	.5
Responded to survey	2,753	100.0	71	2.6	131	4.8	2,539	92.2	12	.4
Provide family planning	578	100.0	6	1.0	23	4.0	547	94.7	2	.3
Do not provide family planning	2,175	100.0	65	3.0	108	4.9	1,192	91.6	10	.5
Did not respond to survey	231	100.0	21	9.0	28	12.1	180	77.9	2	1.0

[1] Includes those institutions which have programs whereas their students can also attend courses at another related institutions.

Table 2. Number and percent of colleges and universities providing family planning services, by control and type of institution.

Type of college or university	Control of Institution									
	Total		Public		Private					
					Religious		Independent		State [1]	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total colleges	2,984	100.0	1,447	48.5	787	26.4	744	24.9	6	.2
Responded to survey	2,753	100.0	1,386	50.4	685	24.9	677	24.6	5	.2
Provide family planning	578	100.0	335	57.9	91	16.0	150	26.0	2	.3
Do not provide family planning	2,175	100.0	1,051	48.3	594	27.3	527	24.2	3	.1
Did not respond to survey	231	100.0	61	26.4	102	44.2	67	29.0	1	.4

[1] These are colleges that are privately administered, but publicly funded.

Table 3. Number and percent of colleges and universities providing family planning services, by highest degree offered and type of institution.

Type of college or university	Highest Degree Offered									
	Total		Less than 4 years		B.S. or B.A.		Advanced Degree		Non-Degree	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total colleges	2,984	100.0	1,158	38.8	875	29.3	946	31.7	5	.2
Responded to survey	2,753	100.0	1,088	39.5	783	28.4	878	31.9	4	.2
Provide family planning	578	100.0	106	18.3	151	26.1	321	55.6	0	0
Do not provide family planning	2,175	100.0	982	45.1	632	29.1	557	25.6	4	.2
Did not respond to survey	231	100.0	70	30.3	92	34.8	68	29.5	1	.4

Table 4. Number and percent of colleges and universities providing family planning services, by student enrollment and type of institution.

Type of college or university	Student Enrollment																			
	Total		Under 200		200-499		500-999		1,000-2,499		2,500-4,999		5,000-9,999		10,000-19,999		20,000 or more		Enrollment Unknown	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Total colleges	2,984	100.0	276	9.2	418	14.0	613	20.5	753	25.2	355	11.9	275	9.2	157	5.3	66	2.3	71	2.4
Responded to survey	2,753	100.0	243	8.8	375	13.6	565	20.5	703	25.5	335	12.2	264	9.6	154	5.6	65	2.4	49	1.8
Provide family planning	578	100.0	15	2.6	31	5.4	87	15.0	137	23.7	80	13.8	93	16.1	79	13.7	47	8.1	9	1.6
Do not provide family planning	2,175	100.0	228	10.6	344	15.8	478	22.0	566	26.0	255	11.7	171	7.9	75	3.4	18	.8	40	1.8
Did not respond to survey	231	100.0	33	14.3	43	18.6	48	20.8	50	21.6	20	8.7	11	4.8	3	1.3	1	.4	22	9.5

Table 5. Number and percent of colleges and universities providing family planning services, by HEW Region and type of institution.

Type of college or university	HEW Regions																							
	Total		01[1]		02[2]		03[3]		04[4]		05[5]		06[6]		07[7]		08[8]		09[9]		10[10]		Unassigned[11]	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Total colleges	2,984	100.0	255	8.5	332	11.1	354	11.9	506	17.0	570	19.1	257	8.6	220	7.4	103	3.5	279	9.3	104	3.5	4	.1
Responded to survey	2,753	100.0	229	8.3	297	10.8	329	12.0	467	17.0	520	18.9	237	8.6	209	7.6	102	3.7	257	9.3	102	3.7	4	.2
Provide family planning	578	100.0	54	9.3	72	12.5	61	10.6	85	14.7	99	17.1	40	6.9	44	7.6	23	4.0	75	13.0	25	4.3	0	0
Do not provide family planning	2,175	100.0	175	8.0	225	10.3	268	12.3	382	17.6	421	19.4	197	9.1	165	7.6	79	3.6	182	8.4	77	3.5	4	.2
Did not respond to survey	231	100.0	26	11.3	35	15.2	25	10.8	39	16.9	50	21.6	20	8.7	11	4.8	1	.4	22	9.5	2	.8	0	0

1/ Includes Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont.

2/ Includes New Jersey, New York, Puerto Rico, Virgin Islands.

3/ Includes Delaware, Washington, D. C., Maryland, Pennsylvania, Virginia, West Virginia.

4/ Includes Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee.

5/ Includes Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin.

6/ Includes Arkansas, Louisiana, New Mexico, Oklahoma, Texas.

7/ Includes Iowa, Kansas, Missouri, Nebraska.

8/ Includes Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming.

9/ Includes Arizona, California, Hawaii, Nevada.

10/ Includes Alaska, Idaho, Oregon, Washington.

11/ Includes American Samoa, Canal Zone, Guam, Micronesia.

SOCIAL SECURITY PAYMENTS AND THE AGED HOUSEHOLD EXPENDITURES IN THE U.S.

Teh-wei Hu and Sharif Ghalib, The Pennsylvania State University*

I. Introduction

During the last decade, the United States has witnessed a marked increase in the aged population (65 years old and over) from 17 million in 1960 to 20 million in 1970. The growth of the aged segment of the population has special significance for the Social Security program. Of the 21.7 million Social Security recipients in 1966, 67 percent received benefits because they or their wives were retired workers aged 65 or over. In 1970, the Social Security program transferred income to 85 percent of the households in the aged cohort.

Social Security benefits have increased substantially in recent years. Benefits of about $1,715 were paid to an average recipient in 1972 (Mullineaux, 1973, p. 3). This figure had increased by about 70 percent of the average annual payment in 1965. The most recent increase was enacted in 1972 when benefits were boosted 20 percent in order to cover the cost-ov-living change. In November, 1973, the Congress has again proposed a 6 to 7 percent increase pending the resolution of the bill.

Many empirical studies have utilized household budget data to analyze the relationship between expenditure and income as well as other socioeconomic variables for the U.S. population as a whole. However, in spite of the growing importance, in both absolute and relative terms, of older consumer units in the U.S., there is a relatively paucity of studies, particularly of the household budget type, that deal with the expenditure patterns of aged family units. Moreover, many of these studies, such as those by Fisher (1955), Brady (1955, 1956), Goldstein (1960, 1965, 1966, 1968), and Reinecke (1971), analyze 1950 or 1960 data and are largely descriptive; only Crockett (1963) and Chu (1972) make use of multiple regression techniques to control for various sociodemographic factors in aged household expenditure patterns.

There is, therefore, a need for further research and more rigorous analysis of the household expenditures of the aged by using regression techniques based on more recent data (1969 to 1970). The objective of this paper is to estimate expenditure functions of the aged population by incorporating Social Security payment information in the model so that the impact of the recent major increase in Social Security benefits on various expenditure items can be determined.

*The authors are, respectively, Professor of Economics and Part-time Instructor, The Pennsylvania State University. The authors are grateful for the financial support of the Social Security Administration. This paper is to be included as a part of the report to be submitted to the Social Security Administration, under Grant No. 56073. The opinions expressed are those of the authors and do not necessarily reflect those of the sponsoring agency.

II. Analytical Framework and the Specification of the Model

Economic theory assumes that a household has a preference function and tries to maximize its satisfaction in choosing the optimal commodity mix, subject to the constraint of household income. In addition to the income factor, it is explicitly recognized that various sociodemographic characteristics--e.g., age, sex, and race--affect these preference orderings and thus the expenditure behavior of households. This recognition is based on both theoretical grounds and previous empirical findings. There is also an important econometric reason for taking these characteristics into account if a goal of the study is to estimate income effects (marginal propensities to expend) using regression analysis: because most sociodemographic variables are correlated with income, if the former variables are omitted from the regression equation, the estimated income coefficients will not show its net effect upon expenditures (they would partially reflect the influences of the excluded variables). Specification errors are introduced when relevant explanatory variables are omitted, resulting in biased estimates of income effects, a characteristic of earlier family budget studies. Therefore, the net influence of income, or "pure income effects" can be picked up only if one controls for relevant sociodemographic variables by explicitly introducing them as regressors in the expenditure functions, thus maintaining the ceteris paribus condition. This procedure does not only permit the estimation of pure income effects, but it also allows us to detect and measure the net influence on expenditures of any single characteristic as the remaining ones are effectively held constant.

Since this paper is also interested in examining the possible impact of Social Security payments (Y_1) on family expenditures, the income variable is decomposed into two parts: account is taken of Social Security payments (Y_1) and other income (Y_2). The use of several income components in the expenditures function is analogous to specifications of aggregate consumption functions in studies by Brown (1952), Klein and Goldberger (1955), and Halbrook and Stafford (1971).

The analyses by Reinecke (1971) and our separate empirical results have shown that, even among the aged households, there are differences in spending behavior between the old (65 to 74 years) and the very old (75 and above). Therefore, a dummy variable (D) is introduced to test and measure the differences in levels of expenditure categories, holding other factors constant. To test and measure possible differences in marginal propensity to expend (MPE), a product of income and dummy variable (DY_1) is introduced.

In addition to variables of income source and age classification, two other groups of variables will be included: (1) economic

variables, such as wealth proxies, debt status, and home ownership; and (2) sociodemographic variables such as household size and composition (classified in terms of single or husband and wife), education, sex, race, employment condition (retired or not retired), physical condition (disabled or not disabled) of the head of household, and region and degree of urbanization of the household. Since the data from two years (1969 to 1970) are pooled into one regression equation, a dummy variable for survey year is introduced to account for possible differences in price and economic conditions between years. Thus, the analytical model is specified as follows:

$$E_i = f(Y_1, Y_2, DY_1, DY_2, D, X_1, X_2, \ldots, X_n, U_i)$$

where E_i represents various expenditure categories (food, alcohol, cigarettes, housing, and the sum of these four expenditures and the value of car(s) owned). Due to data limitations, total expenditures and other types of expenditures are not available. Y_1 denotes Social Security payments; Y_2 denotes other sources of income; D denotes the old age group (65-74); $X_1, X_2, \ldots, X_n$ denote other economic and sociodemographic variables; and U_i is an error term.

A linear equation form is used in this study. The justification is that, for the great majority of expenditure categories, the square of Social Security benefits was not found to differ significantly from zero when the quadratic model was run. This implies that the marginal propensities to expend out of Social Security benefits are, by and large, constant (i.e., independent of the level of income) and that saturation levels are not attained; the latter is not surprising in view of the fact that such saturation levels may be reached, at least in principle, only at relatively high income levels while average annual Social Security benefits are relatively small (about $1,500 a year, or 25 percent of total income). Moreover, the correlation coefficient between Social Security benefits (Y_1) and its square (Y_1^2) was found to be exceedingly high (.96). Given the relatively small size of the Social Security subsamples, the consequence was a high degree of multicollinearity that often gave rise to large standard errors of the coefficients Y_1 and Y_1^2.

III. Data Source and Definition of Variables

The specified models are estimated from both the 1960 to 1961 Survey of Consumer Expenditures (CES, by the U.S. Department of Agriculture and the Bureau of Labor Statistics) and the 1968 to 1971 Panel Study of Income Dynamics (Panel data by the Survey Research Center at the University of Michigan). This paper will present the results of Panel data alone. The 1968 Panel data consisted of a cross section of 4,802 families in the United States (excluding Alaska and Hawaii) who were interviewed four times. By 1971, the sample consisted of 4,840 households, about 750 of them having been newly created as adult members of original Panel units splitting off to form their own families. The Panel study generated a unique data set which provides a wealth of information on sociodemographic characteristics of households in addition to detailed income information by sources and costs of earning income. However, these data do not cover all expenditure items of a household; only the major expenditure items such as food, housing, alcohol, cigarettes, and car are included in the survey.

Since the income data referred to the year prior to the interview while the expenditure data pertained to the survey year, and since 1968 data do not separate the Social Security benefits and other income sources, the complete income and expenditure information was reduced to two years--1969 and 1970. By eliminating the nonaged sample and aged sample with no Social Security benefits, a total of 331 households is retained for the analysis, or a total of 662 households on a two-year pooled basis. The introduction of a dummy variable for survey year has accounted for the possible differences between years.

The definitions of dependent and independent variables in the study are as follows:

Dependent Variables

Food expenditures are defined as the value of annual food consumption, which includes food purchases, values of home-produced food, free food, and one-half of restaurant expenditures (in dollars).

Alcohol and cigarette expenditures are measured on an annual basis (in dollars).

Housing expenditures are defined as the rental value of the residence and utility expenses incurred by the household. For homeowners, they are defined as property taxes, property insurance, mortgage interests, utilities, costs of repair, and the opportunity cost, equivalent to 6 percent of the net equity in the house (in dollars).

Car expenditures are defined only as the three-year average (1968 to 1970) value of cars owned by the household. The Panel data do not provide information on car repair costs and gasoline expenses (in dollars).

Partial total expenditure is the sum of the first four expenditure categories (in dollars).

Independent Variables

Net real disposable income other than Social Security benefits is defined as annual money income minus income taxes, work expenses, and Social Security benefits, plus the value of goods and services produced by, or provided to, the family (in dollars).

Social Security benefits (in dollars).

Partial assets are defined as the sum of estimated savings and net equity in home (in dollars).

Mortgage debt is a dummy variable: 1 = has mortgage debt; 0 = otherwise.

Homeownership is a dummy variable: 1 = homeowner; 0 = otherwise.

Level of education is represented by four categories of dummy variables: (1) less than nine grades; (2) nine grades or more, but less than high school graduation; (3) high school graduate; and (4) education beyond high school. Category (4) is omitted in the regression equation.

Household size is a continuous variable (in number of persons).

Household compositions are in dummy variables: (1) single individuals; (2) husband and wife; (3) other types. Category (3) is omitted in the regression equation.

Age of the head of the household is a dummy variable: 1 = age 65 but less than 74; 0 = otherwise.

Employment status is a dummy variable: (1) = retired; 0 = otherwise.

Location of the household is characterized in two ways, by region and by degree of urbanization. Dummy variables are created for three regions: Northeast, North Central, and South. West is omitted in the regression equation. Urbanization is categorized by dummy variables on the basis of location of sampling unit: (1) population greater than 500,000 (2) population between 50,000 and 500,000 (3) between 10,000 and 50,000, and (4) less than 10,000 persons. Number (4) is omitted in the regression equation.

Sex is a dummy variable: 1 = male; 0 = female.

Race is a dummy variable: 1 = white, 0 = otherwise.

Disability is a dummy variable: 1 = a head of the household who has a physical or nervous condition which completely or severely limits one's productive activities; 0 - otherwise.

Welfare is a dummy variable: 1 = household received any in-kind and/or cash public welfare assistance; 0 = otherwise.

Year is a dummy variable for the year 1970. The dummy variable for the year 1969 is omitted in the equation.

Table 1 presents the means and standard deviations of the variables in the Social Security recipients sample. Several interesting points are revealed.

(1) The average size of household of Social Security recipients is 1.85 members. Of these households, about 38 percent have a family of a single person; 45 percent have husband and wife only.

(2) Among the Social Security recipients, 32 percent are retired, 52 percent are disabled, but only about 9 percent are on welfare.

(3) About 82 percent of Social Security recipients are homeowners, and only 12 percent have outstanding mortgages. Therefore, the net equity of their assets including cash savings is about $12,000, which is about twice their average income of $5,449.

(4) Among the Social Security recipients, 65 percent are between ages 65 and 74. The head of the household is male in about 66 percent of the sample and white in about 90 percent of the sample.

(5) Over 50 percent of the Social Security recipients in the sample have less than a ninth-grade education. About 50 percent of the sample is located in cities with less than 50,000 people. Seventy percent of the sample is from the North Central and Southern parts of the U.S.

(6) Social Security benefits accounted for about 27 percent of the total income for Social Security recipients. They spend most on household items and utilities (30 percent) and food (18 percent). The sum of their alcohol and cigarettes expenditures is only about 1 percent of their total income.

IV. Empirical Estimation

It is well known that cross sectional data on household income and expenditures are subject to errors in measurement. As a result, estimates of MPE will be biased. As Friedman (1957), Summers (1959), and Liviatan (1961) point out, an alternative to overcome these biased estimates is to treat the income as errors in variables and to use either instrumental variables or the two-stage least squares technique to obtain consistent estimates from the model. A commonly used method is to use total expenditures as an instrumental variable for measured income. Since the Panel data do not have information on total expenditures, however, a sum of four separate expenditures (partial total expenditures) is used as an instrumental variable. The results show that regression coefficients become less significant than the classical least squares method. The coefficient of instrumental variables is also rather difficult to use for policy implications and economic interpretation. Furthermore, this study is interested in various income components in relation to expenditures, and the instrumental variable technique is not suitable for this purpose. Therefore, measured incomes are used in the analysis.

Tables of regression results are available from the authors upon request. Table 2 shows the estimated marginal propensity to expend (MPE) of the old aged group and the very old group of the Social Security recipients. Several interesting findings may be summarized with regard to Social Security benefits and income of other sources.

(1) The classification of income into two components (Social Security benefits and income from other sources) is statistically significant to explain the variation of most expenditure items except cigarettes. Apparently, cigarettes are a habit-forming consumption item that is not significantly affected by income.

(2) All MPE's are statistically significant at least at the 10 percent level, except for cigarettes, alcohol (for Social Security benefits, 65-74 age group), and car (for Social Security benefits, 75 and over age group).

(3) For alcohol, housing, and food expenditures, the MPE of Social Security benefits is higher than the MPE of income from other sources. For cigarettes, however, there is no significant difference between the MPE of Social Security benefits and the MPE of income from other sources. The MPE of other income on cars is higher than the MPE of Social Security benefits.

(4) Statistical tests show that the differences of MPE between Social Security benefits and income from other sources are statistically significant within each age group (old and very old), respectively. However, the MPE's of Social Security benefits between the two age groups and the MPE's of income from other sources between the two age groups are not statistically significant. Thus, it is concluded that the classification of age group (old and very old) is not as important as the classification of

TABLE 1

Means and Standard Deviations of Variables: Social Security Subsample Pooled; 1969-1970

Variables	Means and Standard Deviations	Variables	Means and Standard Deviations
Soc Sec B (Y_1)	$1494.28 (721.35)[a]	Male	.657 (.475)
Husband/Wife	.452 (.-98)	Homeowner	.822 (.383
Single	.375 (.484	White	.900 (.300)
Retired C	.323 (.468)	City GE 500	.174 (.379)
Disabled	.518 (.500)	City GE 50	.233 (.423)
Mortgage Debt	.125 (.331)	City GE 10	.307 (.461)
Household Size	1.85 (1.06)	Northeast	.177 (.382)
Ed LE 8	.545 (.498)	Northcentral	.326 (.469)
Ed Le 9-11	.156 (.363)	South	.369 (.483)
Ed Le H.S.	.086 (.281)	Alcohol	$35.50 (114.64)
Welfare	.092 (.289)	Cigarette	$33.86 (76.15)
Pt Asst	$11,885.22 (11,793.85)	Hset a Util	$1,633.66 (1,111.57)
ODRY (Y_2)	$3,955.44 (4,444.34)	AFoodC	$987.88 (583.06)
Year = 70	.530 (.499)	Av Car V	$537.83 (755.82)
Age 65 to 74 D	.650 (.477)	Pt Exp	$2,690.92 (1,431.07)
Sample Size	662		

[a]Standard deviations are in parentheses

Social Security benefits and income from other sources. The dummy variable for age is also not statistically significant for every expenditure category.

(5) For the very old group (75 and above), the magnitude of MPE of Social Security benefits is highest for food, followed by housing. There is no influence on cigarettes and car. The negative MPE on alcohol is not a plausible result. However, the magnitudes of MPE of income from other sources is highest for car, followed by food, alcohol, and housing.

(6) For the old age group (between 65-74), the magnitude of MPE of Social Security benefits is highest for car, followed by housing and food. There is no influence on cigarettes and alcohol. The magnitude of MPE of income from other sources is highest for car, followed by food and alcohol.

Other findings with regard to sociodemographic factors and expenditure relations are as follows:

(1) Household size has a significant positive effect on food expenditures but not on other items. On the other hand, single-person families spend much less on most items than other types of families.

(2) The head of the household who is disabled spends significantly less on food and car than other types of families.

(3) Other things being equal, welfare

TABLE 2

Marginal Propensities to Expend: Social Security Sample, 1969-1970

MPE of Income Sources for Aged Group \ Expenditures		Alcohol	Cigarettes	Housing and Utilities	Food	Average Car(s) Value	Partial Expenditures
Age 75+	MPE (Social Security)	-.021**	-.012	.076**	.128***	-.020	.171**
	MPE (Other Income)	.020**	.002	.011*	.024***	.061***	.057***
Age 65-74	MPE (Social Security)	.010	.003	.092***	.088***	.114**	.193***
	MPE (Other Income)	.006***	.001	.006	.036***	.087***	.049**

Notes: *** significant at the 1% level (two-tailed test).
** significant at the 5% level (two-tailed test).
* significant at the 10% level (two-tailed test).

families of the aged group spend more on food and cigarettes but less on housing and car than non-welfare families.

(4) For the Social Security recipient sample, there is no significant difference in expenditure patterns whether the head of the household is male or female, or whether the head is retired or otherwise. However, the head of the household who is white spends more on cigarettes, housing, and food than the heads of nonwhite households.

(5) Households that are located in large cities with population over 500,000 spend more on all expenditure categories except car (they spend less on the value of their car).

(6) Households that are located in the North Central region spend less on food and housing but more on car than households in the West. There is no significant difference among regions with respect to alcohol and cigarette expenditures.

V. Policy Implications and Concluding Remarks

The empirical results presented in the previous section show that aged Social Security recipients have a higher marginal propensity to expend on food and housing out of Social Security benefits than the marginal propensity expended from income of other sources. This finding implies that additional increases in Social Security benefits will most likely result in an increase on food and housing rather than on other non-necessity items such as alcohol and cigarettes. Two reasons may explain the above findings: (1) Social Security benefits are considered by recipients as a constant source of income; therefore, this amount of income is allocated for the necessities; and (2) Social Security recipients, on the average, have a much lower level of income ($5,445, including the Social Security benefits) as compared to an average income of $8,300 for all households in the Panel data. In fact, the average income of the Social Security recipients is even lower than the average income of welfare families in the sample ($5,700) as shown by Hu and Knaub (1973). Furthermore, the Social Security benefits account for about 25 percent of the household's total income. The budget allowance on food and housing is about 50 percent of the income. Thus, it is most likely that the Social Security recipient will spend the payment more on food and housing than on any other items.

One of the objectives of this study is to estimate the impact of the proposed or enacted increases in incomes from Social Security benefits on expenditures of aged recipients. The increase will amount to about $200 for single persons and $400 for couples. The estimated marginal propensity to expend with respect to Social Security benefits, as shown in Table 2, suggests that there will be a significant increase in food and housing expenditures. Suppose the increase of Social Security benefits is about $200 for a typical recipient. Then the very old group (75 and above will spend an extra $26 on food and $15 on housing but no extra amount on alcohol, cigarettes, and car. On the other hand, the old age group (between 65 and 74) will spend an extra $17 on food, $18 on housing, and $23 on car. Given the 1966 information on the number of people who receive Social Security (14.5 million), the additional increase of $200 per recipient will generate at least an extra $290 million on food and $217 million on housing expenditures (assuming each recipient spends an extra $20, on the average, for food and $15 for housing). Therefore, the demand for food and housing items will increase as a result of an increase in Social Security benefits.

References

Brady, Dorthy S. "Family Saving, 1888-1950," in R. W. Goldsmith, D. S. Brady and H. Menderhausen, A Study of Saving in the United States, Vol. III, New Jersey: Princeton University Press, 1956.

Brady, Dorothy S. "Influence of Age on Saving and Spending Patterns," Monthly Labor Review Vol. 78 (November, 1955, No. 2), 1240-44.

Brown, T. M. "Habit Persistence and Lags in Consumer Behavior," Econometrica, Vol. 20 (July, 1952, No. 3), 355-371.

Chu, Kwang-wen. "Consumption Patterns Among Different Age Groups," Unpublished Ph. D. Dissertation, University of California, Los Angeles, 1972.

Crockett, Jean. "Older People as Consumers," Aging and the Economy, edited by H. L. Orbach and Clark Tibbits, Ann Arbor: University of Michigan Press, 1963.

Fisher, Janet A. "Family Life Cycle Analysis in Research on Consumer Behavior," Consumer Behavior, Vol. II: The Life Cycle and Consumer Behavior, edited by Lincoln H. Clark, New York: New York University Press, 1955.

_____. "Income, Spending and Saving Patterns of Consumer Units in Different Age Groups," Studies in Income and Wealth, Vol. 15. New York: National Bureau of Economic Research, 1952.

Friedman, Milton. A Theory of the Consumption Function, New Jersey: Princeton University Press, 1957.

Goldstein, Sidney. "Changing Income and Consumption Patterns of the Aged, 1950-1960," Journal of Gerontology, Vol. 20 (October, 1965, No. 4), 453-461.

_____. Consumption Patterns of the Aged, Philadelphia: University of Pennsylvania Press, 1960.

_____. "Home Tenure and Expenditure Patterns of the Aged, 1960-61," Long-Range Program and Research Needs in Aging and Related Fields, Hearings before a Senate Special Committee on Aging, Part I: C.C.H., 1968, 268-278.

Goldstein, Sidney. "Urban and Rural Differentials in Consumer Patterns of the Aged, 1960-61," Rural Sociology, Vol. 31 (September, 1966, No. 3), 333-345.

Holbrook, Robert and Stafford, Frank. "The Propensity to Consume Separate Types of Income: A Generalized Permanent Income Hypothesis," Econometrica, Vol. 39 (January, 1971, No. 1), 1-21.

Klein, L. R. and Goldberger, A. S. An Econometric Model of the United States, 1929-1952, Amsterdam: North-Holland Publishing Co., 1955.

Liviatan, Nissan. "Errors in Variables and Engel Curve Analysis," Econometrica, Vol. 29 (July, 1961, No. 3), 336-362.

Morgan, James and Smith, James. A Panel Study of Income Dynamics, Ann Arbor, Michigan: Survey Research Center, University of Michigan, 1971.

Mullineaux, Donald J. "Paying for Social Security: Is it Time to 'Retire' the Payroll Tax?" Business Review, Federal Reserve Bank of Philadelphia, April, 1973, 3-10.

Summers, R. "A Note on Least-Squares Bias in Household Expenditure Analysis," Econometrica, Vol. 27 (January, 1959, No. 1), 121-26.

U.S. Department of Health, Education and Welfare. Social Security Administration, Office of Research and Statistics, Expenditures of Two-Person Units and Individuals After Age 55, by John A. Reinecke, Staff Paper No. 9, Washington, D.C.: Government Printing Office, 1971.

COMBINING TWO DICHOTOMOUS RESPONSES CONTAINING RESPONSE ERRORS

Her Tzai Huang, Grambling College
Wayne A. Fuller, Iowa State University

Introduction

In sample surveys, response errors often constitute a sizeable portion of the total error associated with an estimator. Hansen, et al. [7] presented a mathematical model for survey observations containing response errors. Response errors in a binomial population were studied by Hansen, et al. [6]. They emphasized the difference in the impact of uncorrelated and correlated response deviations on the sampling properties of estimators. Since then a number of papers have been devoted to the effects of misclassification on estimates and tests associated with multinomial problems (e.g. [1], [5], [10], [12]).

Given the presence of response errors a questionnaire containing two responses for the variable of interest is sometimes used. For example, "How old are you?" and "When were you born?" can be used to obtain two (possibly different) responses for the age of sample individuals. The objective of such a questionnaire is to obtain a value for each individual somehow superior to which can be obtained from a single question. The presence of two questions requires a rule for combining the two answers.

There are a number of references dealing with the problem of combining estimators of a common mean if the sampling is from a normal distribution. If we have a number of estimates $y_i (i = 1,2,\ldots,k)$ normally and independently distributed about the same mean, μ, with known variances σ_i^2, the minimum variance unbiased estimator of μ is the weighted mean, $\bar{y}_w = \sum_{i=1}^{k} W_i y_i / W$, where $W_i = \sigma_i^{-2}$ and $W = \Sigma W_i$. When the σ_i^2 are unknown, they may be replaced by their unbiased estimators, s_i^2. Properties of such estimators have been investigated by Cochran [2], Meier [11], Cochran and Carroll [3], Huang [9] and others.

In this paper, we consider a finite universe in which individuals are classified into one of two classes, "1" or "0." The proportion of individuals in class 1 is denoted by P, and the proportion of individuals in class 0 denoted by Q, where $P + Q = 1$. We assume that a simple random sample of size n is drawn from this universe, and each individual responds to two questions that permit him to be classified into one of the two groups. These may be two questions on a single questionnaire or they could be questions on two different questionnaires. Due to the response errors, the two responses are not always the same. We consider the estimation of the population proportion P and the classification of sample individuals into the two classes. It is assumed that a super-population of responses exists for each individual. Let

- p_{mu} denote the probability that an individual who belongs to class 1 answers 1 to the m-th question;
- q_{mu} denote the probability that an individual who belongs to class 1 answers 0 to the m-th question;
- p_{mv} denote the probability that an individual who belongs to class 0 answers 1 to the m-th question;
- q_{mv} denote the probability that an individual who belongs to class 0 answers 0 to the m-th question.

We denote the response to the m-th question by sample individual i by Y_{mi} and assume that the response probabilities are such that the sample responses are unbiased for the population proportions, i.e.,

$$E(Y_{mi}) = p_{mu}P + p_{mv}Q = P, \qquad (1)$$

where the expectation is over individuals and responses.

Classification of Individuals Given a Third 0-1 Variable

In the continuous variable case it is possible to use the sample information to estimate weights by which the two responses may be combined (see Huang [9]). However, in the classification case additional information beyond that contained in the two responses seems to be required for efficient combination. We assume that a third zero-one variable, X_3, is available from the questionnaire. We also assume i) X_3 has nonzero correlation with the individual true value and ii) the response errors in Y are independent of X_3. Note that X_3 may contain response error provided that the response error is independent of that in Y.

Each individual response can be identified by one of eight 3-tuples, $Z_i = (Y_{1i}, Y_{2i}, X_{3i}) =$ (0, 0, 0), (1, 0, 0), (0, 1, 0), (1, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1), (1, 1, 1). We identify these eight possibilities by a single subscript, $j = 1,2,\ldots,8$ and let

- R denote the probability that X_{3i} equals 1;
- α denote the conditional probability that the true value for individual i, $\mu_{.i}$ equals 1 given that X_{3i} is 1;
- β denote the conditional probability that $\mu_{.i}$ equals 1 given that X_{3i} is 0;
- P_j denote the conditional probability that $\mu_{.i}$ equals 1 given case j, $j = 1,2,\ldots,8$;

$P_{(j)}$ denote the j-th conditional probability in the ordered arrangement of the P_j's,

$$P_{(1)} \le P_{(2)} \le \cdots \le P_{(8)},\ j = 1,2,\ldots,8.$$

$$P_{j_1j_2j_3} = P(Y_{1i} = j_1,\ Y_{2i} = j_2,\ X_{3i} = j_3),$$
$$j_1 = 0,1;\ j_2 = 0,1;\ j_3 = 0,1;$$

$F_{(j)} = P_{j_1j_2j_3}$ where the $F_{(j)}$ are ordered by the magnitude of P_j, $j = 1,2,\ldots,8$.

It is easily seen that $\beta(1-R) + \alpha R = P$ and the probabilities $P_{j_1j_2j_3}$ are given by

$$P_{000} = q_{1v}q_{2v}(1-\beta)(1-R) + q_{1u}q_{2u}\beta(1-R) \quad (2.1)$$

$$P_{100} = p_{1v}q_{2v}(1-\beta)(1-R) + p_{1u}q_{2u}\beta(1-R) \quad (2.2)$$

$$P_{010} = q_{1v}p_{2v}(1-\beta)(1-R) + q_{1u}p_{2u}\beta(1-R) \quad (2.3)$$

$$P_{110} = p_{1v}p_{2v}(1-\beta)(1-R) + p_{1u}p_{2u}\beta(1-R) \quad (2.4)$$

$$P_{001} = q_{1v}q_{2v}(1-\alpha)R + q_{1u}q_{2u}\alpha R \quad (2.5)$$

$$P_{101} = p_{1v}q_{2v}(1-\alpha)R + p_{1u}q_{2u}\alpha R \quad (2.6)$$

$$P_{011} = q_{1v}p_{2v}(1-\alpha)R + q_{1u}p_{2u}\alpha R \quad (2.7)$$

$$P_{111} = p_{1v}p_{2v}(1-\alpha)R + p_{1u}p_{2u}\alpha R \quad . \quad (2.8)$$

Further

$$P_1 = \frac{q_{1u}q_{2u}\beta}{q_{1v}q_{2v}(1-\beta) + q_{1u}q_{2u}\beta} \quad (3.1)$$

$$P_2 = \frac{p_{1u}q_{2u}\beta}{p_{1v}q_{2v}(1-\beta) + p_{1u}q_{2u}\beta} \quad (3.2)$$

$$P_3 = \frac{q_{1u}p_{2u}\beta}{q_{1v}p_{2v}(1-\beta) + q_{1u}p_{2u}\beta} \quad (3.3)$$

$$P_4 = \frac{p_{1u}p_{2u}\beta}{p_{1v}p_{2v}(1-\beta) + p_{1u}p_{2u}\beta} \quad (3.4)$$

$$P_5 = \frac{q_{1u}q_{2u}\alpha}{q_{1v}q_{2v}(1-\alpha) + q_{1u}q_{2u}\alpha} \quad (3.5)$$

$$P_6 = \frac{p_{1u}q_{2u}\alpha}{p_{1v}q_{2v}(1-\alpha) + p_{1u}q_{2u}\alpha} \quad (3.6)$$

$$P_7 = \frac{q_{1u}p_{2u}\alpha}{q_{1v}p_{2v}(1-\alpha) + q_{1u}p_{2u}\alpha} \quad (3.7)$$

$$P_8 = \frac{p_{1u}p_{2u}\alpha}{p_{1v}p_{2v}(1-\alpha) + p_{1u}p_{2u}\alpha} \quad (3.8)$$

We first develop a rule for classifying individuals assuming the population parameters known. We wish an assigned value for each individual, $\hat{\mu}_{.i}$, where $\hat{\mu}_{.i}$ is either zero or one and $\hat{\mu}_{.i}$ is chosen to minimize the expected squared error loss

$$E\{(\hat{\mu}_{.i} - \mu_{.i})^2\}$$

subject to the restriction that the mean of $\hat{\mu}_{.i}$ is an unbiased estimator of P,

$$E\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{.i}\} = P \quad . \quad (4)$$

We propose the following rule:

Rule: Order the P_j's such that

$$P_{(8)} \ge P_{(7)} \cdots \ge P_{(2)} \ge P_{(1)},$$

and define c to be the index such that

$$\sum_{j=c+1}^{8} F_{(j)} < P \text{ and } \sum_{j=c}^{8} F_{(j)} \ge P \quad .$$

Assign

$$\hat{\mu}_{.i(j)} = 1, \quad j=c+1, c+2, \ldots, 8$$

$$\hat{\mu}_{.i(c)} = 1, \text{ with probability A}$$

$$= 0, \text{ with probability 1-A}$$

$$\hat{\mu}_{.i(j)} = 0, \quad j=1,2,\ldots,c-1$$

where

$$A = \frac{1}{F_{(c)}}[P - F_{(8)} - F_{(7)} - \cdots - F_{(c+1)}] \quad . \quad (5)$$

Theorem: The Rule minimizes

$$E\{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{.i} - \mu_{.i})^2\}$$

subject to

$$E\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{.i}\} = P \quad .$$

Proof: The proof is by induction. We first show that a randomized rule applied to any additional case will result in a larger mean square error than our rule. We then assume that our rule is better than randomizing any r cases and show that it is also better than randomizing any (r+1) cases. The details are contained in Huang [9].

The average mean square error of classification is given by

$$E\{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{.i} - \mu_{.i})^2\} = \sum_{j=1}^{c-1} F_{(j)}P_{(j)} + F_{(c)}[A + P_{(c)} - 2AP_{(c)}] + \sum_{j=c+1}^{8} F_{(j)}[1-P_{(j)}] \quad (6)$$

Estimation of Parameters

In the practical situation the parameters of interest must be estimated from the sample data. We denote the eight sample proportions by

$$\hat{P}_{000}, \hat{P}_{100}, \hat{P}_{010}, \hat{P}_{110}, \hat{P}_{001}, \hat{P}_{101}, \hat{P}_{011}, \hat{P}_{111} . \quad (7)$$

There are five independent parameters, say p_{1u}, p_{2u}, P, α, R, the remaining parameters being defined by identities and the unbiasedness restrictions. Define $\theta' = (p_{1u}, p_{2u}, P, \alpha, R)$, $f(\theta) = (P_{000}, P_{100}, P_{010}, P_{110}, P_{001}, P_{101}, P_{011})'$ and $Y = (\hat{P}_{000}, \hat{P}_{100}, \hat{P}_{010}, \hat{P}_{110}, \hat{P}_{001}, \hat{P}_{101}, \hat{P}_{011})'$ then we may express the observed proportions as

$$Y = f(\theta) + e , \quad (8)$$

where $E\{e\} = 0$. The covariance matrix of e is that of the multinomial with parameters $f(\theta)$. The Gauss-Newton method of estimation (see Fuller [4] and Hartley [8]) may then be used to solve this non-linear regression problem.

Example

The Statistical Laboratory of Iowa State University in cooperation with the Statistical Reporting Service of the U.S. Department of Agriculture conducted a survey of 262 Iowa farm operators in September and October of 1970. In both of these interviews the respondents were asked to name the most important product of their farm operation. A good deal of information on the farm operation was also collected. We consider the variables

Y_{mi} = 1 if the i^{th} farm operator reports hogs the most important product on the m^{th} interview, m = 1,2
= 0 otherwise

X_{3i} = 1 if the number of breeding hogs is equal to or greater than 30 for the i^{th} farm operator,
= 0 otherwise.

The analysis is summarized in Table 1. The estimated parameters obtained by the Gauss-Newton procedure are $(\tilde{p}_{1u}, \tilde{p}_{2u}, \tilde{P}, \tilde{\alpha}, \tilde{R})$ = (0.9153, 0.9159, 0.3890, 0.6716, 0.3585). The estimated standard errors for these estimators are (0.0357, 0.0356, 0.0281, 0.0498, 0.0295).

In this particular example the two methods of obtaining the information are two identical questions asked at different times. Therefore we would expect the values of p_{1u} and p_{2u} to be about the same. The estimates for these parameters are approximately equal.

Using the estimates the conditional probability that the true value is 1 is estimated for each cell and is given on the fourth line of the table. On the basis of these estimated conditional probabilities we assign the estimated individual true value, $\tilde{\mu}_{.i}$, as follows:

$\tilde{\mu}_{.i}$ = 1 if individual i belongs to case (1,1,1), (1,1,0), (0,1,1) or (1,0,1),

= 1 for a random sample of 4 of the 13 individuals who belong to the case (0,1,0),

= 0 otherwise.

The estimated mean square error of this classification, $\hat{MSE}(\tilde{\mu}_{.i})$, is 0.0439. If we use only the first question the estimated classification error is 0.0659 and if we use the second question alone the estimated classification error is 0.0654. Thus the use of two questions and the auxiliary information has reduced the estimated classification error by about one third.

Acknowledgement

This research was partially supported by Joint Statistical Agreement J.S.A. 73-2 with the Bureau of the Census, U. S. Department of Commerce.

REFERENCES

[1] Bryson, M. R., "Errors of classification in a binomial population," Journal of the American Statistical Association 60 (1965), 217-224.

[2] Cochran, W. G., "Problems arising in the analysis of a series of experiments," Journal of the Royal Statistical Society Supplement 4 (1937), 102-118.

[3] Cochran, W. G. and Carroll, S. P., "A sampling investigation of the efficiency of weighting inversely as the estimated variance," Biometrics 9 (1953), 447-459.

[4] Fuller, W. A., "Gauss-Newton estimation with a preliminary estimate," Unpublished class notes, Department of Statistics, Iowa State University, 1972.

[5] Giesbrecht, F. G., "Classification errors and measures of association in contingency tables," Proceedings of the Social Statistics Section of the American Statistical Association 1967, 271-276.

[6] Hansen, M. H., Hurwitz, W. N. and Bershad, M. A., "Measurement errors in censuses and surveys," Bulletin de Institute International de Statistique 38, No. 2 (1961), 359-374.

[7] Hansen, M. H., Hurwitz, W. N., Marks, E. S. and Mauldin, W. P., "Response errors in surveys," Journal of the American Statistical Association 46 (1951), 147-190.

[8] Hartley, H. O., "The Gauss-Newton method for the fitting of non-linear regression functions by least squares,: Technometrics 3 (1961), 269-280.

[9] Huang, H. T., "Combining multiple responses in sample surveys," Unpublished Ph.D. dissertation, Iowa State University, 1972.

[10] Koch, G. G., "The effect of non-sampling errors on measures of association in 2 x 2 contingency tables," Journal of the American Statistical Association 64 (1969), 852-863.

[11] Meier, P., "Variance of a weighting mean," Biometrics 9 (1953), 59-73.

[12] Mote, V. L. and Anderson, R. L., "An investigation of the effects of misclassification on the properties of χ^2-tests in the analysis of categorical data," Biometrika 52 (1965), 95-109.

Table 1. Most Important Product (Example)

(Y_{1i}, Y_{2i}, X_{3i})	(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)	Total
Obs. Frequency	116	8	13	32	29	3	8	53	262
Obs. Proportion	0.4428	0.0305	0.0496	0.1221	0.1107	0.0115	0.0305	0.2023	1.0000
Est. Model Prob.	0.4428	0.0366	0.0365	0.1257	0.1071	0.0245	0.0246	0.2022	1.0000
Est. Cond. Prob. that $\mu = 1$	0.0024	0.3120	0.3154	0.9887	0.0160	0.7552	0.7581	0.9983	-
$\tilde{\mu}_{.i}$	0	0	0.3288*	1	0	1	1	1	-

*Randomization probability

FERTILITY DECLINE IN CANADA, 1961 - 1970: AN ANALYSIS THROUGH FERTILITY TABLE APPROACH

Ashraf K. Kayani, Alberta Department of Health
and Social Development
P. Krishnan, University of Alberta

Social Scientists have evinced keen interest in the fertility decline in the North American Continent. Studies carried out in the U.S.A. (Ryder and Westoff, 1971), and Canada (Allingham, Balakrishnan, and Kantner, 1970) reveal the important role played by the pill in the decline of fertility during the late sixties. The decrease has been so significant that the U.S.A. fertility has already headed toward a ZPG. A recent press release claimed that Canada also is moving in a ZPG direction. The conclusion arrived at, for the press release, is based on the measure of total fertility. The present writers feel that the fertility decline as deduced from the changing total fertility rates is defective on account of methodological problems associated with the conventional measures of fertility. Furthermore, Canada has a large flow of immigrants and this also may influence the fertility rates. This paper attempts to remedy a major methodological problem of the conventional measures of natality by means of the fertility table technique. The different demographic components of Canadian fertility decline are also presented.

2. Methodological Problem of Conventional Fertility Measures

A rate is defined as the ratio of the frequency of occurrence of an event during the period (usually one year) under consideration to the appropriate population at risk. The computation of proper risk population is not always easy in view of the continuous nature of the occurrence of vital events. In mortality studies, proper exposure for the population under consideration is introduced. The present writers are of the opinion that the different conventional measures do not take cognisance of the correct woman years of exposure.

The core of the measurement problem lies in calculating the woman years of exposure. In the building of stochastic process models of fertility, the non-fertility period when a woman is pregnant, is taken care of (Dandekar, 1955; Singh, 1963, 1964, 1968, Singh & Bhattacharya, 1970. We believe that the non-fertile period has to be considered when any fertility measure is computed. All the females who give birth in one year may not beget children in the next year on account of biomedical reasons. While some bring forth issues, some others may not be exposed at all to the risk of child-bearing. Therefore, during the next year the woman years of exposure are bound to be less than those of the initial year assuming that there are no new entrants to the child-bearing span by means of immigration.

The importance of the point made here is further clarified by the following analysis. We make three assumptions.

1. A female who has been now delivered of a (live-birth) child would not conceive (beget) another child for at least two months after and shall be infertile for another 9 months if she conceives.
2. All live births are uniformally distributed over the 12 months of a year.
3. Females in the reproduction ages do not die. This "mortality-free" assumption can be dropped when mortality factor is incorporated into the fertility table idea proposed here.

Assumptions 1 and 2 are the crucial elements for the determination of the risk population. It has been estimated by medical researchers that a normal live birth is the outcome of 38 - 42 weeks of pregnancy. Also a female who has just given birth to a child, on an average, would not be able to conceive sometimes on account of bio-medical reasons (post partum amenorrhoea).

Consider year Z. The females who are delivered of babies in the first two months of Z may give birth to a child in any month of year Z + 1. In other words, they are exposed to the risk of bringing forth a child for 12 months as far as year Z + 1 is concerned. But the females who are responsible for live births in the months of March through December of year Z are exposed in year Z + 1 for 11, 10, 9, 8, 7, 6, 5, 4, 3 and 2 months respectively. Hence the total months of exposure to mothers of live birth's occurring in Z as far as year Z + 1 is concerned would be equal to

$$12 + 12 + 11 + 10 + \ldots\ldots + 2 = 89.$$

Months of non-exposure = 144 - 89 = 55

Thus it is clear that 55/144, that is 38.2 per cent of mothers in year Z are, on an average, not exposed to the risk of bringing forth a child in year Z + 1. The "reduction" to the risk population is considerable and has to be given its due. In section 3, we suggest the use of the life table approach as a solution to this methodological handicap.

3. Fertility Tables

Fertility tables, which are analogous to the life tables, are a way out of this methodological problem. If birth spacing data are available, the fertile life tables suggested by Pool and Wright (1969) can be fruitfully employed.

The following are the main functions of a fertility table:

q_x - Probability of a live-birth in year (x, x + 4) at the beginning of the age interval.

l_x - Number of females at risk of giving birth to at the beginning of the age interval (x, x + 4).

L_x - A first approximation of woman-years of exposure to the risk of a child birth in (x, x +4).

T_x - $\sum_{X=15}^{45} L_x$

E_x - T_x/l_x A first approximation of the expected years of exposure to the risk of child

birth in the reproductive span.

Y_x - The number of years the females who have births to are not at risk of giving birth to another child in (x, x + 4).

$Lx' = L_x - Y_x$ - A better approximation of the woman-years of exposure to the risk of child in (x, x + 4).

$T'_x = \sum_{X=15}^{45} (L_x - Y_x)$

$E'_x = T'_x / l_x$ - A better approximation of the expected years of exposure to the risk of child birth in the reprodctive span years lost due to child birth at age X.

$E_x - E'_x$ - Years lost due to child birth at age x.

T.F.R., or expected family size at age x --- $(E_x - E'_x)$ 12/11.

4. Fertility Tables for Canada 1961 - 1970

Fertility tables for the 1961 - 70 decade are presented in appendix (tables A - 1 to A - 5). The age-specific fertility rates are used as probabilities; no conversion of M_x's into q_x's as done if life table construction was performed. The synthetic and the period TFR's for the decade under review are shown in Table 1.

4.1 Virtual Fertility Decline

The decennial decline in TFR as measured by Statistics Canada estimates is 1,529 births per 1,000 woman, while the fertility table approach reveals a decline of 1,464 births. Thus for every 1,000 women in the reproductive span overestimates of 65 births over the decade is noted by Statistics Canada by employing a conventional measure of fertility. It is worthwhile to point out that, from 1961 through 1964, the fertility table values are lower than the period TFR, in 1965 they are almost identical, and from 1966 onwards the period rates are lower than the fertility table ones. Since the fertility table rates are synthetic values, they look like cohort rates. The convergence and divergence of period and cohort rates have been noted for Canada (Henripin, 1972). It seems to the present authors that, however synthetic, the fertility table rates convey the behavior of cohort rates. A detailed study of this problem is being undertaken.

4.2 Immigration Component

Immigrant females of reproductive age-groups in Canada constitute 1 to 3 per cent of the total Canadian females of reproductive ages.

The impact of immigration on Canadian fertility needs to be probed into. Economic insecurity, problems related with assimilation in the host community etc. all lead to postponement of births. We suggest two models to study the effect of immigration on the conventionally computed TFR.

Assumption 1: Immigrant females in the reproductive age group are not likely to give birth to a child in the first year of this stay in Canada.

Assumption 2: Immigrant females in the reproductive age group are not likely to give birth to a child in the first two years of their stay in Canada.

From the data available on immigration by age and month of arrival (Statistics Canada publications Vital Statistics 1961 through 1970), one can develop estimates of the TFR under assumptions 1 and 2. These are presented in Table 1. The first type of assumption yields a TFR of 3,874 in 1961 and 2,344 in 1970 for the non-immigrant Canadian population. The ten-year decline works out at 1,530 births per 1,000 women which is not far apart from that given by the difference in period TFR values. Under assumption 2, the decline in 1962 - 70 is 1,444 births per 1,000 women, showing that the immigrant component leads to a superficial reduction of 85 births per 1,000 women. The realistic situation may be somewhere between those referred to by assumptions 1 and 2. Hence, on an average, a superficial reduction of about 43 births per 1,000 women is the effect of immigration on Canadian fertility decline during the period 1961 - 70.

Conclusion

This analysis of fertility decline done here for Canada reveals that the overall decrease in the decade 1961 - 70 as portrayed by changing TFR is exaggerated by about 65 births per 1,000 women. If immigration component is built into the study, the superficiality of the decline is further increased. Overestimates of total fertility rate are in direct conflict with the claims by some that Canada is close enough to achieve zero population growth.

Acknowledgements

Computations by Colleen Kimak are gratefully acknowledged.

References

J.D. Allingham, T.R. Balakishnan, and J.F. Kantner (1970), "The End of a Rapid Increase in the Use of Oral Anovulants", Demography 7, 31 - 42.

V.M. Dandekar (1955), "Certain Modified Forms of Bronomial and Poisson Distributors", Sankhya, 15, 237 - 250.

J. Henripin (1972), Trends and Factors of Fertility in Canada. 1961 Census Monograph. Ottawa: Information Canada.

D.I. Pool, and M. Wright (1969), La Calculation de Disperance de Fertile an Ghana", Paper presented at the meetings of the Canadian African Studies Association.

N.B. Ryder, and C.W. Westoff (1971), Reproduction in the United States, 1965, Princeton: Princeton University Press.

S.N. Singh (1963), "Probability Models for the Variation in the Number of Births for Couple", Journal of American Statistical Association, 158, 721 - 727.

S.N. Singh (1964), "A Probability Model for Couple

Fertility", Sankhya Series B, 26, 89 - 94.

S.N. Singh (1968), "A Chance Mechanism of Variation in Number of Births for Couple", Journal of American Statistical Association, 63, 209 - 43.

S.N. Singh and B.N. Bhattacharya (1970), "A Generalized Probability Distribution for Couple Fertility", Brometrics, 26, 33 - 40.

Data Sources

Statistics Canada. (1972) Vital Statistics 1970.

OTTAWA: Information Canada

APPENDIX

Due to the limitations of space, fertility tables for 1961, 1963, 1965, 1967 and 1970 are selected.

Table 1 Total Fertility Rates Canada: 1961 - 1970

YEAR	TFR STAT CANADA	FERTILITY TABLE TFR	TFR IMMIGRATION 1	UNDER ASSUMPTION 2
1961	3840	3801	3874	
1962	3756	3721	3791	3827
1963	3669	3639	3710	3745
1964	3502	3482	3544	3583
1965	3145	3146	3193	3232
1966	2812	2827	2867	2911
1967	2586	2606	2646	2698
1968	2441	2466	2488	2546
1969	2388	2412	2428	2522
1970	2311	2337	2344	2383

Source: Vital Statistics, 1970 Ottawa: Information Canada.
FERTILITY TABLES

Table A1 Fertility Table for Canada Females 1961

AGE GROUP	qx	lx	Lx	Tx	Ex	L'x	T'x	E'x	T.F.R.
15 - 19	.058200	100000.0	491723.9	3369349.	33.69	466683.	3052567.	30.526	3.801
20 - 24	.233600	97941.6	466379.2	2877625.	28.78	371051.	2585884.	25.859	3.501
25 - 29	.219200	92220.6	462913.4	2411245.	24.11	374127.	2214832.	22.148	2.357
30 - 34	.144900	92665.1	472893.2	1948331.	19.48	412936.	1840705.	18.407	1.292
35 - 39	.081100	95027.7	483707.0	1475437.	14.75	449382.	1427768.	14.278	0.572
40 - 44	.028500	97154.7	493098.5	991730.	9.92	480802.	978386.	9.784	0.160
45 - 49	.002400	98981.3	498635.6	498631.	4.99	497588.	497584.	4.976	0.013

Table A2 Fertility Table for Canada Females 1963

AGE GROUP	qx	lx	Lx	Tx	Ex	L_x'	T_x'	E_x'	T.F.R.
15 - 19	.053100	100000.0	492438.9	3374912.	33.75	469559.	3071636.	30.716	3.639
20 - 24	.226000	98118.6	467509.9	2882473.	28.82	375060.	2602077.	26.021	3.365
25 - 29	.210600	92454.6	464218.2	2414963.	24.15	378674.	2227017.	22.270	2.255
30 - 34	.140300	92932.5	473752.4	1950744.	19.51	415593.	1848342.	18.483	1.229
35 - 39	.075800	95177.9	484578.2	1476991.	14.77	452439.	1432748.	14.327	0.531
40 - 44	.025900	97335.7	493645.2	992413.	9.92	482458.	980309.	9.803	0.145
45 - 49	.002100	99073.4	498770.7	498768.	4.99	497854.	497851.	4.979	0.011

Table A3 Fertility Table for Canada Females 1965

AGE GROUP	qx	lx	Lx	Tx	Ex	L_x'	T_x'	E_x'	T.F.R.
15 - 19	.049300	100000.0	492972.7	3391858.	33.92	471707.	3129656.	31.297	3.146
20 - 24	.188600	98250.8	472450.6	2898885.	28.99	394485.	2657948.	26.579	2.891
25 - 29	.181900	93623.7	468981.7	2426434.	24.26	394338.	226346.3	22.635	1.956
30 - 34	.119400	93836.3	477386.3	1957452.	19.57	427511.	1869125.	18.691	1.060
35 - 39	.065900	95866.6	486610.9	1480065.	14.80	458552.	1441613.	14.416	0.461
40 - 44	.022000	97675.6	494534.9	993454.	9.93	485015.	983061.	9.831	0.125
45 - 49	.002000	99211.8	498923.6	498919.	4.99	498050.	498046.	4.980	0.010

Table A4 Fertility Table for Canada Females 1967

AGE GROUP	qx	lx	Lx	Tx	Ex	L'x	T'x	E'x	T.F.R.
15 - 19	.045200	100000.0	493550.1	3410413.	34.10	474030.	3193214.	31.932	2.606
20 - 24	.161100	98394.0	476189.4	2916862.	29.17	409065.	2719183.	27.192	2.372
25 - 29	.151400	95402.3	473783.3	2440672.	24.41	411019.	2310118.	23.101	1.567
30 - 34	.091400	94816.2	482099.2	1966888.	19.67	443543.	1899099.	18.991	0.813
35 - 39	.050600	96804.9	489652.2	1484788.	14.85	467973.	1455555.	14.556	0.351
40 - 44	.015900	98205.6	495928.9	995136.	9.95	489029.	987582.	9.876	0.091
45 - 49	.001500	99429.1	499212.6	499207.	4.99	498557.	498553.	4.986	0.008

Table A5 Fertility Table for Canada Females 1970

AGE GROUP	qx	lx	Lx	Tx	Ex	L'x	T'x	E'x	T.F.R.
15 - 19	.43400	100000.0	493803.9	3419672.	34.20	475052.	3224959.	32.250	2.337
20 - 24	.142100	98456.9	478770.3	2925868.	29.26	419241.	2749907.	27.499	2.112
25 - 29	.145600	95119.1	475132.7	2447097.	24.47	414601.	2330665.	23.307	1.397
30 - 34	.080700	95004.9	483739.4	1971964.	19.72	449581.	1916064.	19.161	0.671
35 - 39	.038500	97168.3	491703.4	1488224.	14.88	475139.	1466482.	14.665	0.261
40 - 44	.011000	98628.8	497049.9	996521.	9.97	492266.	991343.	9.913	0.062
45 - 49	.000900	99604.3	499474.4	499471.	4.99	499081.	499077.	4.991	0.005

STAGES AND DETERMINANTS OF THE STUDENT'S DECISION-MAKING PROCESS IN THE CHOICE OF AN EDUCATIONAL INSTITUTION

Arno Kleimenhagen, University of Wisconsin-Whitewater
G. M. Naidu, University of Wisconsin-Whitewater

Higher education in the United States has been experiencing a dramatic change for the past five years. Every administrator now recognizes that he is operating in a "buyers' market" instead of a "sellers' market." With declining enrollments, budget squeezes, and the resulting financial crisis in higher education, student recruitment is receiving increased attention from faculty and university administrators. Central to this issue is the understanding of how students make decisions in the choice of a campus. It is the purpose of this paper to define the stages and identify the relevant variables that influence each stage of the student's decision-making process.

THE EDUCATIONAL BUYING PROCESS MODEL

The model discussed is a special case application of the standard consumer buying process model found in the marketing literature. [1], [2] The student's choice of an institution of higher education is similar to the brand choice of a consumer in a business environment. A typical student, graduating from a high school or college, goes through five distinct stages in his decision-making process of selecting a campus. These stages include problem recognition (felt need for higher education), institution comprehension (search), evaluation of alternatives, enrollment, and postenrollment feelings as depicted in Figure 1. [2]

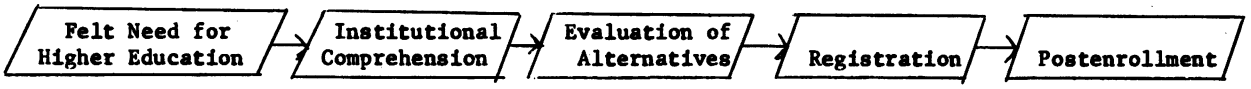


FIG. 1. Stages of Student's Educational Buying Process

The decision to enroll in an institution is illustrated in Figure 2.

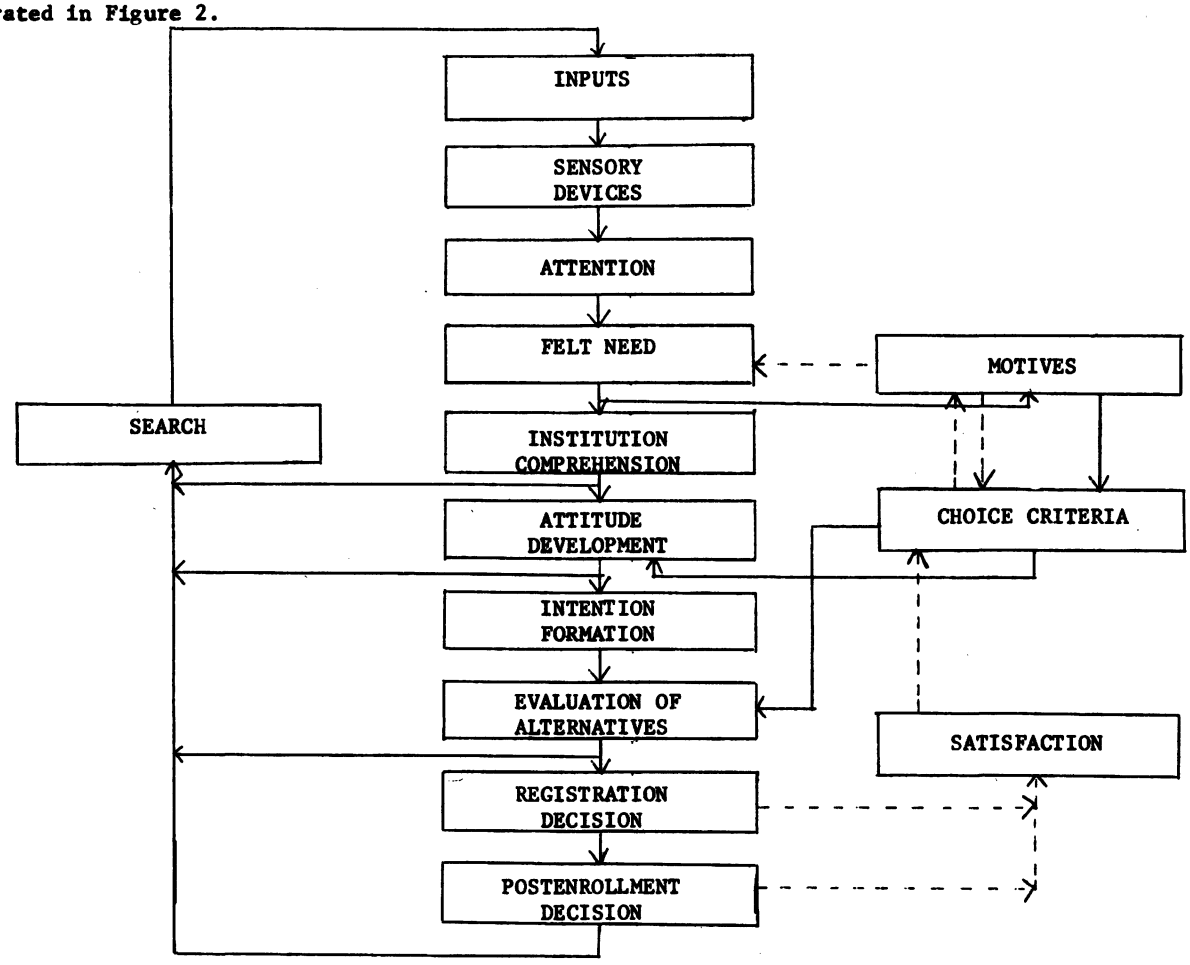


FIG. 2. Decision to Enroll at a Particular Campus

Felt Need

In the case of higher education, the first stage involves the process whereby the student becomes aware of his or her need to pursue additional study. This first stage like the subsequent stages is complex, involving the interactions of many variables, including motives, attitudes, values, reference group influences and societal pressures. Edward McDill and James Coleman conclude in their study of the Illinois high school graduates that parental pressures and the peer group pressures are the most significant variables in influencing the primary demand for higher education. [3] In a similar study, Joseph Katz concludes that societal pressure is the most important variable influencing the primary demand. [4] The question of "why" people attend college has been analyzed by Iffert on the basis of a national sample survey of students in twenty institutions. He concludes that a "better paying job" and "compelling interest in a particular field of study" are the two main reasons for going to college.[5] Some other influential variables at this stage include: dissatisfaction with present occupational opportunities as obtainable through his or her existing level of education, changed reference groups, and changed financial status.

From a marketing point of view the more students who recognize the need for higher education, the greater will be the numbers who ultimately enroll. Economically, primary demand for higher education is generated at this stage. Society (through appropriate governmental agencies) and the institutions of higher education themselves should promote the primary need by calling attention to the benefits of higher learning. The sophisticated approach is not through paid advertising,but rather through publicity and personal selling. Personal selling as applied to higher education refers to the professional activities of the faculty in the areas of research and public service.

Institution Comprehension

Once the felt need has been recognized and the student decides to pursue additional education, the process involves the student's ability to comprehend the unique qualities and educational opportunities of different institutions. This includes two sub-stages--one concerning "what to buy?" (program or vocational choice) and two, "where to enroll?"

What educational program to pursue is very much dependent on the student's aptitude and self-actualization goals, and perceived economic rewards. [6] Friedman and Kuznets indicate that as the income from a profession increases the number of graduates in the field also increases with some time lag. [7] In some cases, the student may select an institution with vague career plans and may keep his options open to a later time period.

The stage of a student's decision making process as to which college to apply is dependent on two factors--his "evoked set" and his "evaluative criteria." [8] His evoked set is the set of all institutions that the student is aware of and is interested in out of the total population of institutions of higher learning. The evoked set is influenced by geographical location, peers, counselors, family socio-economic status, communications through mass media, athletic programs, and the student's aptitude, occupational, and academic goals. [6]

To which college he will apply is influenced by his "evaluative criteria." This includes the school's image, faculty reputation, program recognition, tuition, and other costs, campus and off-campus living facilities, part-time or full-time job opportunities on or off campus, physical facilities like computers, libraries, and student unions, general campus appearance, and social life on the campus. [6] The student may obtain information on the above variables through a relative, a friend or classmate, or by going through the formal channels of writing for information to a selected list of institutions.

The evaluative criteria varies from student to student. A student with high scholastic aptitude may give greater importance to the school's reputation or faculty recognition whereas an average student may have "price" or "location" as the most significant variable.

Using his evaluative criteria, the student formalizes within himself a set of ordered preferences based on his perception of the institutional images and applies for admission. To keep his options open, he applies to more than one institution while awaiting acceptance from his "top choice" institution. [7]

At this stage, the promotional activities of a typical institution rely primarily on personal visits by the director of admissions to various high schools in order to generate interest in their programs to the prospective candidates. There is evidence to believe that the students are becoming more selective and that competition among institutions is increasing as the percentage of "shows" on most campuses have been falling steadily. Good tactics to follow at this stage include promptness in sending requested information and fast processing of applications. Some institutions even grant "instant admissions" in order to gain favorable impressions from prospective students. Favorable institutional images can be obtained through mass media; personal selling (visits of campus administrators and faculty to various feeder schools, etc.) and other channels of communication like direct mail and display booths at shopping centers can generate more inquiries from potentially interested student populations.

The appropriate general institutional strategy directed to this stage is to reinforce the perceptions of prospective students who favor your institution and to invoke a perceived risk factor in those who do not include your institution as one of his or her alternative choices. This,

from a marketing point of view, is very tricky business. You run the danger of modifying favorable predispositions toward your institution without necessarily modifying the converse situation.

Institution-dominated sources of information (advertising, catalogs, bulletins, etc.) are important but probably do not accomplish much more than reinforcement. Favorable institutional image, professionally recognized faculty, and program strength no doubt carry the most weight in influencing behavior during this process. Additional institutional strategy for this stage should be directed at achieving program accreditation and promoting faculty research and other professional activities making certain these are publicized.

The Evaluation of Alternatives

When the student learns which institutions have admitted him, he then reaches the point in the decision-making process where he chooses a campus. The variables that influence this stage are again dependent on the scholastic quality of the student as well as his previous level of educational attainment. [6], [9] In the Naidu study the following hypotheses were generated:

i) The higher the level of a student, the greater the relative importance of faculty reputation in the choice of a university.

ii) The lower the level, the greater the relative importance placed on general reputation.

iii) The higher the level, the greater the relative importance of financial aid.

iv) The higher the level, the lower the relative importance of the location of the campus.

v) The higher the level, the lower the relative importance of appearance of campus.

vi) The lower the level, the greater the relative importance of costs (including tuition.

It is important to recognize that in evaluating the evoked set of alternatives, a typical student looks at a total package of value satisfactions. It is the sum of perceived importance of each of the variables based on evaluation criteria and the rating of those variables that determine final choice. These variables which make up the "package" include the general reputation of the institution, faculty reputation, program reputation, the quality of instruction, the quality of physical facilities, costs, individual attention, area job opportunities for the student and/or his/her spouse, and the campus location among others. [10] It is at this stage that most institution-dominated information sources are directed. Effective tactics for this segment of the decision process include:

(1) Efficient and prompt processing of admission applications.

(2) Campus visits - interviews with faculty.

(3) A thorough follow-up program to all who express some interest in attending your campus.

(4) Thorough counseling for all prospects.

In some cases it is the timing of the offer which may be extremely important. In a study conducted at Michigan State University, Allan Grimes points out that "poor timing, inefficient handling of materials, ambiguous communication and assignment of advisors not in their field" were some of the major reasons for the no shows (MSU). [11] A sound administrative policy to increase the percentage of shows at this stage is to offer the "right package" to the "right student" at the "right time." This calls for very good coordination within various campus offices such as the admissions office, financial aids, housing, etc.

Enrollment

Stage four (enrollment) deals with the educational environment interaction. Two possible outcomes could occur--attendance or "halt." [1] The process may halt because no alternative satisfies or exposure to the educational environment may alter the relationship between the desired and actual state. Effective tactics used in stage three should continue through the registration period of the student into an appropriate academic program. Marketing success at this stage involves the blending of institution characteristics to meet the educational needs of the market. Administrators and faculty must work together at achieving this product-market accord.

Postenrollment

The final stage deals with the postenrollment behavior of students. The institutional goals are retention and graduation. Strategies closely tied to this stage involve curriculum development, adequate academic counseling programs, exposure to major societal, political, and social issues, alumni relations, community interaction and so on. The usefulness or value of the received education is ultimately assessed in the market place when the student graduates. The value in the market place is dependent on the creativity of the faculty in their area of research, public service and curriculum and the administration for providing such a productive organizational climate.

Even though the immediate rewards go to the stu-

dents, in the long run both the educational institution and society are beneficiaries. With successful alumni, the reputation of the institution is enhanced. The community will recognize the fine efforts of the institution. Above all, this has a synergetic effect in attracting more and better quality faculty. A study at MSU indicates that the faculty choice of a campus is very much influenced by the quality of incoming students. [12] Obviously the quality of institutional output is dependent on the quality of inputs.

CONCLUSIONS

In order to effectively recruit students and thereby offset the financial ravages brought on with declining enrollments, it is incumbent for university administrators to better understand the student's decision-making process in the choice of an educational institution. It is hoped that through conceptualization of this type, progress will be made toward a comprehensive understanding of the students' decision-making process in selecting an institution or program of higher education. It is further hoped that this understanding will improve and direct relevant academic programming.

REFERENCES

[1] Engel, Kollet and Blackwell, Consumer Behavior (New York: Holt, Rinehart & Winston, 1968), pp. 347 - 519.

[2] Philip Kotler, Advanced Marketing Management: Analysis, Planning, and Control (New York: Prentice-Hall, 1967), p. 68.

[3] Edward L. McDill and James Coleman, "Family and Peer Influence in College Plans of High School Students," Sociology of Education, Winter, 1965, Vol. 38, No. 2.

[4] Joseph Katz, "Societal Expectations and Influences," The College and the Student, ed. by L. Dennis and J. Kauffman (Washington: American Council on Education, 1966).

[5] Robert E. Iffert, "Retention and Withdrawal of College Students," U.S. Dept. of Health, Education and Welfare, Bulletin No. 1, 1958.

[6] Gurramkonda M. Naidu, Systems Approach to Marketing Aspects of Higher Education (Unpublished Ph.D. Thesis, Michigan State University, 1969).

[7] Milton Friedman and S. Kuznets, Income from Independent Professional Practices, National Bureau of Economic Research, New York, publication No. 45, 1945).

[8] John A. Howard and Jagadish N. Sheth, The Theory of Buyer Behavior (New York: John Wiley & Sons, Inc., 1969).

[9] Gurramkonda M. Naidu, "Marketing Strategies for Higher Education," Proceedings of the Fall Conference, the American Marketing Association, 1970.

[10] Gurramkonda M. Naidu and Donald G. Leeseberg, Determinants of College Choice: A Case Study," Proceedings of the Fourth Annual Meeting of the American Institute for Decision Sciences, 1972.

[11] Allan Grimes, "Survey of No-Shows," Memorandum, Dept. of Political Science, Michigan State University, January, 1968.

[12] Edwin F. Commack, "Faculty Mobility and Productivity and Achievement at Michigan State University," (Unpublished Manuscript, Michigan State University, 1964).

AN EXPERIMENT WITH ALTERNATE RESPONDENT RULES IN THE NATIONAL HEALTH INTERVIEW SURVEY

Mary Grace Kovar and Robert A. Wright
National Center for Health Statistics

The program of the National Center for Health Statistics includes a variety of data collection systems designed to assemble information on the health of the United States population. One system in this program is the National Health Interview Survey (H.I.S.), a continuous nationwide sample survey of households in which household members are interviewed by Bureau of the Census interviewers to obtain information about illness, disability, medical care, and other health related items.

In assessing the reliability of the statistics derived from the H.I.S. a major concern has been the effect on the data of a respondent rule which permits any adult family member to report for other family members. Studies which have been conducted by H.I.S. to evaluate the validity of specific types of information obtained in the H.I.S., studies which involved comparison of information obtained in the interview with that obtained from records, have indicated that the most important problem in household interviews is underreporting, and that the degree of underreporting tends to be more severe when the information is obtained through a proxy rather than from the person himself.[1]

The possibility of other respondents reporting less illness and disability than the person would have reported for himself has been recognized for many years. The Hunterdon County Health Survey found that persons reporting for themselves "reported proportionately nearly half again as many more of the conditions found on subsequent clinical examinations as when persons were reported for by other family members."[2] Results from the Baltimore Health Survey were much the same.[3]

In the California Health Survey, reinterview data indicated that proxy respondents reported less illness than self-respondents, and that respondents other than the spouse were primarily responsible for the large net differences between self and proxy respondents.[4]

In studies where an attempt was made to control whether the individual was a self-respondent or had a proxy, results were not consistent. In a 90-household pilot study conducted by the London School of Hygiene and Tropical Medicine, more illnesses were reported when all adults were self-respondents than when wives reported for the entire family. This was particularly true for males.[5] In the report on males age 35 or older from the North Dakota study, no significant differences were found.[6] The data from the Charlotte Pretest of the H.I.S. also indicated that in households where all adults were required to be self-respondents, reported illness and disability rates were higher than in households where related adults could report for those not at home when the interviewer called, but sampling and response variability were too high to permit definitive conclusions.[7]

"As a result of this lack of conclusiveness in the evidence available, the extra cost of interviewing all adults for themselves was not considered a good investment."[7] Therefore, since its inception in 1957 the H.I.S. has used a respondent rule which states that adults at home when the interviewer calls should respond for themselves, but the information for all children under age 17, for adults who are incapable of being interviewed, and for adults not at home at the time of the interview is to be obtained from another adult family member such as a parent or spouse. However, the validation studies referred to above, which were not designed to test the effect of self or proxy respondents, have indicated that further research was needed.

Consequently, as part of the continuing research program of H.I.S., a special study designed to measure the degree to which the use of proxy respondents affects the national statistics was conducted during the second quarter (April, May and June) of 1972.[8] This paper is concerned primarily with the field implementation of the study and a brief report on the substantive data.

The respondent rule study was carried out as part of the ongoing National Health Interview Survey. Including an experiment as part of an ongoing survey is desirable when the purpose is to test the effect of modifying specific procedures within the framework of the operating data collection system. It does, however, impose certain limitations on the study design as the integrity of the statistics from the ongoing survey should not be compromised.

The study required a control sample of households to be interviewed using the standard respondent rule and an experimental sample using a self-respondent rule. For the experimental households every adult capable of being a self-respondent was required to be one. Proxy respondents were still used for all children under age 17, but were accepted for adults only if the adult was incapable of responding because of disability (usually senility) or absence over the entire interview period.

Interviewing under alternate procedures had to go on throughout the quarter to control for changes in health conditions and utilization due to seasonality. Several alternative study designs, including randomizing the interviewers' assignments, were considered and then rejected for administrative reasons. The design finally adopted was to assign pairs of weeks to each rule so that all households scheduled for interviewing during the first week of the thirteen week quarter were interviewed under the standard rule, during the next two weeks under the self-respondent rule, during the following two weeks under the standard

rule, and so on through the quarter. This paired week design was adopted on the recommendation of the Bureau of the Census field supervisors to minimize confusion on the part of the interviewers and to make office administration of the study as simple as possible. The design, the number of households in the sample, the number of households interviewed, and the number of persons in the interviewed households are shown in Table 1.

The only other change in field procedures was that the rules for overtime were relaxed so that the interviewers could make all the additional call-backs needed to interview each adult for himself.

The interviewers were instructed to carry out their assignments during the experimental periods exactly as they did during the ongoing survey. They were particularly instructed not to shift their initial calls to later in the day in hope of finding more adults at home and thus easing their burden and reducing the number of call-backs.

Such a shift would mean that data from the experimental weeks on persons who would have had proxy respondents would not be comparable to the data from the control weeks on persons who did have proxy respondents. We needed these data to evaluate whether the difference between self and proxy respondents which we had found in the survey was due to reporting errors or to actual differences in the population.

The key question was whether the interviewers made more initial contacts in the evening (6 PM or later) during the experimental weeks than during the control weeks. There is evidence that some interviewers did deviate from their usual procedures particularly in suburban areas. During the experimental weeks approximately 22 percent of the calls were made at 6 PM or later, compared with 19 percent during the control weeks. The comparable percentages in suburban areas are 23 and 18 percent.

As a result, the proportion of adults not at home at the time of the initial call was lower during the experimental than during the control weeks. During the experimental weeks only 26 percent of the adults were not at home and thus required an additional call to convert them from proxy to self-respondents. During the control weeks, 37 percent of the adults were not at home and had proxy respondents. The National estimates based on the experimental and control weeks are not affected by this deviation from design specifications, but evaluation of field costs and some of the comparisons between procedures are affected.

The introduction of a self-respondent rule was expected to introduce some new problems in data collection and to increase some existing ones. In general, there were fewer problems than anticipated.

Response rates were almost unaffected. The household response rate remained unchanged at 96 percent. The individual response rate (persons within interviewed households for whom information was obtained) decreased only from 99.85 percent during the control weeks to 98.72 percent during the experimental weeks. Obviously response rates were still high enough that problems of response would not preclude adopting a self-respondent rule.

Interviewing schedules were maintained remarkably well. During the control weeks almost 94 percent and during the experimental weeks 93 percent of the interviews were completed during the scheduled week. During the control weeks 68 percent and during the experimental weeks 79 percent of the interviews were actually completed during the first three days of the scheduled week (Table 2).

The cost of utilizing a self-respondent rule in household interviews was the third major concern. We measured cost by: number of calls required to complete the interview, monetary cost of the field work, and interviewers' subjective judgement.

During the experimental weeks the average number of calls was 2.53 per household--36 percent more than the 1.86 calls per household during the control weeks (Table 3).

The percentage increase in the average number of calls was approximately the same for households located in urban (34 percent), suburban (36 percent), or rural non-farm (38 percent) areas. Rural farm households required 47 percent more calls during the experimental weeks. Only 3 percent of the households were in rural farm areas but, because of travel time and distance, the large increase in the number of calls could have a disproportionate effect on monetary costs.

The monetary cost of introducing the self-respondent rule was calculated by the Bureau of the Census which kept special records of field costs for both the experimental and control weeks. Overall, the nonlisting costs for the experimental weeks were about 17 percent higher than they would have been without the self-respondent rule. The 95 percent confidence interval around the 17 percent cost increase is from 5 to 28 percent.

These measures of cost increase are upper limits. The interviewers were working under the field instructions designed to keep the experimental and control weeks comparable, additional calls and overtime were authorized, and additional record keeping was required. These inefficient procedures, which were instituted as part of the study design, increased the cost beyond what would be expected in the ongoing Health Interview Survey which would utilize more efficient methods.

The interviewers' subjective evaluation was that any improvement in the quality of the data under this particular self-respondent rule was not worth the cost of collecting it. The necessity for more evening calls presented major problems for them--particularly in urban areas. HIS interviewers are women and they did not feel safe interviewing at night. Several said that if the rule became part of the survey, they would be forced to quit; others said that "It took the fun out of interviewing." They were willing to carry out the experiment but the prospect of implementing the self-respondent rule permanently would cause them to reevaluate their participation.

A higher rate of interviewer turnover is a cost factor which we cannot measure. On the other hand, it takes 18 months for an interviewer to reach peak efficiency so we certainly cannot ignore the impact of increased turnover either on costs or on the quality of the data.[9]

The experiment was successful in demonstrating that it is possible to institute a self-respondent rule in a national survey if you are willing to pay for it. In contrast to the control weeks when 67 percent of the adults aged 19 or older were self-respondents, 96 percent were self-respondents during the experimental weeks. As shown in Table 4, the great difference was for males; instead of 49 percent there were 95 percent self-respondents. An unexpected difference was that during the experimental weeks mothers were more likely to respond for children under 17, particularly girls, than during the control weeks. This difference for the children is noteworthy as the interviewers repeatedly stated that the rule change they would like to see is a tightening in the rule about who is eligible to respond for children.

The experiment was also successful in detecting differences in health measures based on the two rules. Our hypothesis had been that rates of illness, disability, and outpatient utilization based on a self-respondent rule would be higher than those based on the standard rule. Of the ten routinely collected measures which we analyzed six were significantly higher under the self-respondent rule using a one-tailed test (Table 5). We find this impressive as these are relatively objective measures, items which are not subject to large respondent bias as more subjective measures such as attitudes are, and because sampling errors based on six weeks of data collection are large, particularly for the two-week recall items, which makes it difficult to detect differences.

We have tried to predict the effect a self-respondent rule would have on H.I.S. estimates. The results are given in Table 6. If a self-respondent rule had been in effect in 1971, we might have estimated 225 million more days of restricted activity, 123 million more doctor visits and 2.4 million more persons who were limited in their usual activity. The confidence intervals around these estimates are large and a much larger sample would be needed to speak with any confidence.

In the future, the question of deciding who is an eligible respondent is expected to become more critical as the National Health Interview Survey moves into questions on attitude, costs of health care, extent of insurance coverage and other areas where personal or specialized knowledge is being elicited. Fortunately, the routine items are not so sensitive but the experiment has demonstrated that it is possible, if necessary, to collect information directly from the household member best qualified to give it. For children that is the person responsible for their care, usually the mother, and for most adults it is the individual himself.

References

[1]National Center for Health Statistics, Vital and Health Statistics, PHS Pub. 1000, Series 2, Public Health Service, Washington, D.C., U.S. Government Printing Office: "Measurement of Personal Health Expenditures," No. 2, (June 1963); "Health Interview Responses Compared With Medical Records," No. 7, (July 1965); "Comparison of Hospitalization Reporting in Three Survey Procedures," No. 8, (July 1965); "Interview Responses on Health Insurance Compared With Insurance Records," No. 18 (August 1966); "Interview Data on Chronic Conditions Compared With Information Derived From Medical Records," No. 23, (May 1967); "Development and Evaluation of an Expanded Hearing Loss Scale Questionnaire," No. 37, (April 1970); "Optimum Recall Period for Reporting Persons Injured in Motor Vehicle Accidents," No. 50, (April 1972); "Net Differences in Interview Data on Chronic Conditions and Information Derived from Medical Records," No. 57, (June 1973).

[2]Elinson, Jack, and Trussell, Ray E., "Some Factors Relating to Degree of Correspondence for Diagnostic Information as Obtained by Household Interview and Clinical Examinations," American Journal of Public Health, 49, (March 1957), 311-321.

[3]Krueger, Dean E., "Measurement of Prevalence of Chronic Disease by Household Interviews and Clinical Evaluation," American Journal of Public Health, 47, (1957) 953-960.

[4]Mooney, H. William, "Methodology in Two California Health Surveys," Public Health Monograph No. 70, PHS Pub. No. 942 (1962).

[5]Cartwright, Ann, "The Effect of Obtaining Information from Different Informants on a Family Morbidity Inquiry," Applied Statistics, 6 (1), (March 1957), 18-25.

[6]Enterline, Philip E., and Capt, Katherine G., "A Validation of Information Provided by Household Respondents in Health Surveys," American Journal of Public Health, 49(2), (February 1959), 205-212.

[7]Nisselson, Harold, and Woolsey, Theodore D., "Some Problems of the Household Interview Design for the National Health Survey," Journal of the American Statistical Association, 54, (March 1959), 69-87.

[8]Haase, Kenneth W., and Wilson, Ronald W., "The Study Design of an Experiment to Measure the Effects of Using Proxy Responses in the National Health Interview Survey," Proceedings of the American Statistical Association, Social Statistics Section, (1972), 289-293.

[9]National Center for Health Statistics, Vital and Health Statistics, DHEW Pub. (HSM) 73-1328, Series 2, Public Health Service, Washington, D.C., U.S. Government Printing Office, "Quality Control and Measurement of Nonsampling Error in the Health Interview Survey," No. 54 (March 1973).

Table 1. Experimental Design, Number of Households in Sample and With Completed Interviews, and Number of Persons in Interviewed Households: National Health Interview Survey, United States, April-June, 1972.

Week	Week Designation	Respondent Rule	Number of		
			Households in Sample	Households Interviewed	Persons in Interviewed Households
1	Control	Standard	881	850	2,643
2	Experimental	Self	918	895	2,556
3	Experimental	Self	903	879	2,708
4	Control	Standard	829	806	2,483
5	Control	Standard	880	857	2,574
6	Experimental	Self	912	873	2,477
7	Experimental	Self	841	806	2,313
8	Control	Standard	931	897	2,778
9	Control	Standard	877	841	2,497
10	Experimental	Self	888	843	2,495
11	Experimental	Self	924	873	2,629
12	Control	Standard	916	867	2,693
13	Control	Standard	866	820	2,477
Total	All Weeks		11,566	11,107	33,323
	Control	Standard	6,180	5,938	18,145
	Experimental	Self	5,386	5,169	15,178

Table 2. Percent Distribution of Completed Interviews According to Day and Week of Completion for Control and Experimental Weeks: National Health Interview Survey, United States, April-June, 1972.

Week Interview Completed / Day	Percent Distribution of Completed Interviews	
	Control Weeks	Experimental Weeks
Total	100.0	100.0
Scheduled Week	93.8	93.0
Monday	20.0	21.7
Tuesday	25.4	27.2
Wednesday	22.9	21.3
Thursday	14.7	12.8
Friday	6.1	5.6
Saturday	4.4	4.0
Sunday	0.2	0.5
Second Week	4.8	5.1
Monday	2.4	2.2
Tuesday	1.0	1.6
Wedn.-Sunday	1.4	1.4
Third Week or Later	0.3	0.9
Not Ascertained	1.1	1.0

Table 3. Average Number of Calls Required to Complete Interviews in Sample Households by Area of Residence and by Time of Day of Initial Contact for Control and Experimental Weeks: National Health Interview Survey, United States, April-June, 1972.

Residence / Time of Initial Contact	Average Number of Calls per Household	
	Control Weeks	Experimental Weeks
All Residence Areas	1.86	2.53
Inside SMSA	1.97	2.66
Central City	2.04	2.74
Outside Central City	1.90	2.58
Outside SMSA	1.65	2.28
Nonfarm	1.67	2.30
Farm	1.46	2.15
All Times	1.86	2.53
Before Noon	1.75	2.38
Noon - 6 PM	1.70	2.40
6 PM or Later	2.46	3.00
Time Not Recorded	1.00	2.39

Table 4. Percent Distribution of Persons According to Who Responded by Age and Sex of Sample Person for Control and Experimental Weeks: National Health Interview Survey, United States, April-June, 1972.

Age Sex Week Designation	Total	Respondent				
		Self	Spouse	Mother	Father	Other and Unknown
Age: Under 17 Years						
Both Sexes						
Control	100.0	0.3	0.0	84.4	11.1	4.2
Experimental	100.0	0.3	-	87.2	9.5	3.0
Males						
Control	100.0	0.0	0.0	85.1	11.2	3.6
Experimental	100.0	0.0	0.0	86.8	10.5	2.7
Females						
Control	100.0	0.5	0.1	83.7	10.9	4.8
Experimental	100.0	0.6	-	87.6	8.6	3.3
Age: 17-18 Years						
Both Sexes						
Control	100.0	24.5	1.2	57.7	10.0	6.7
Experimental	100.0	23.7	0.8	63.1	9.1	3.3
Males						
Control	100.0	17.2	1.3	62.2	11.3	8.1
Experimental	100.0	15.6	1.3	69.0	9.4	4.6
Females						
Control	100.0	31.9	0.9	53.2	8.6	5.4
Experimental	100.0	31.3	0.3	57.5	8.8	2.1
Age: 19-44 Years						
Both Sexes						
Control	100.0	63.7	24.0	8.2	1.7	2.3
Experimental	100.0	96.6	1.7	0.6	0.2	0.9
Males						
Control	100.0	43.5	40.8	10.7	2.3	2.8
Experimental	100.0	95.5	2.7	0.7	0.3	0.9
Females						
Control	100.0	82.6	8.3	5.9	1.2	1.9
Experimental	100.0	97.4	0.9	0.6	0.1	1.0
Age: 45-64 Years						
Both Sexes						
Control	100.0	67.5	27.3	0.7	0.0	4.5
Experimental	100.0	96.5	2.5	0.2	-	0.7
Males						
Control	100.0	48.4	46.0	0.8	0.1	4.7
Experimental	100.0	95.4	3.7	0.2	-	0.7
Females						
Control	100.0	84.6	10.5	0.6	0.0	4.2
Experimental	100.0	97.6	1.5	0.2	-	0.7
Age: 65 Years or Older						
Both Sexes						
Control	100.0	80.6	12.2	0.2	-	7.1
Experimental	100.0	93.7	2.5	-	-	3.8
Males						
Control	100.0	71.7	23.1	-	-	5.2
Experimental	100.0	93.9	4.2	-	-	1.9
Females						
Control	100.0	86.9	4.3	0.3	-	8.4
Experimental	100.0	93.7	1.2	-	-	5.1

Table 5. Rates for Selected Health Measures According to Respondent Rule Used and the Percent Difference by Type of Recall Question: National Health Interview Survey, United States, April-June, 1972.

Type of Recall Question / Health Measure	Respondent Rule		Percent Difference[1]	Z Statistic
	Self	Standard		
Two-Week Recall	Rate per 100 persons per quarter			
Restricted Activity Days	404.3	377.4	7.1	1.173
Bed Days	141.1	148.9	-5.2	.686
Work-Loss Days	140.7	117.6	19.6	1.694
Doctor Visits	128.9	114.8	12.3	1.696
Dental Visits	36.4	38.3	-5.0	.587
Acute Conditions	47.9	42.6	12.4	1.698
Six-Month Recall	Rate per 100 persons per year			
Hospital Discharges	14.7	13.8	6.5	1.198
Twelve-Month Recall or Prevalence	Percentage of persons with			
Limitation of Activity	13.6	12.4	9.7	2.838
Limitation of Mobility	3.6	3.1	16.1	1.745
Doctor Visits in 12 Months	73.6	72.0	2.2	21.560

[1]Percent Difference = $\left(\frac{\text{Self - Standard}}{\text{Standard}}\right) \times 100$

Table 6. National Estimates of Selected Health Measures for 1971 and Estimates of Change Under Self-Respondent Rules by Type of Recall Question: National Health Interview Survey, United States, 1971 and April-June, 1972.

Type of Recall Question / Health Measure	National Estimate 1971	Change Under Self-Respondent Rule		
		Estimate	Confidence Intervals	
			One Standard Deviation	Two Standard Deviations
	(in thousands)			
Two-Week Recall				
Restricted Activity Days	3,175,594	225,467	192,263	384,526
Bed Days	1,238,873	-64,421	93,924	187,849
Work-Loss Days	396,210	77,657	45,882	91,764
Doctor Visits	999,289	122,913	72,506	145,012
Dental Visits	311,943	-15,597	26,578	53,156
Acute Conditions	442,203	54,833	32,311	64,622
Six-Month Recall				
Hospital Discharges	27,571	1,792	1,496	2,992
Twelve-Month Recall or Prevalence				
Limitation of Activity	24,817	2,407	848	1,697
Limitation of Mobility	NA	NA	NA	NA
Doctor Visits in 12 Months	146,465	3,222	150	300

Source 1971 Estimate: Current Estimates from the Health Interview Survey, Series 10, No. 79, DHEW Publication No. (HSM) 73-1505.

ECONOMIC FLUCTUATIONS & INTERNAL MIGRATION IN THE UNITED STATES

G. K. Kripalani, Western Michigan University

INTRODUCTION

Investigations relating to factors influencing internal migration decisions contribute to better understanding of why people move and help to improve judgments about the extent and magnitude of future population adjustments. Such analyses permit insights into the relative importance of causal factors underlying population change.

Several widely divergent motives may underlie the migration behavioral pattern of the people of an area. Better wages, or more generally, more favorable economic opportunities represent one major group of factors influencing migration decisions. The level of economic activity in the whole economy is an important determinant included in this group. Another major group of causes stems from socio-cultural environment of the areas of origin of migrants and their anticipated evaluation of corresponding socio-cultural situations in the areas of potential in-migration. Migration decisions are also affected by information, costs, existence of programs of assistance and kindred factors.

The principal premise that underlies this study is that there are at least a few major independent variables affecting net migration and that some of these are non-measurable or non-observable, and that valid data series for such variables do not exist for use in empirical investigations. The method of analyses used is, therefore, designed to recognize and take into account this problem of nonobservability of some of the major explanatory variables. It is further recognized that net migration behavior patterns vary between the races, between the sexes and between age groups within each race-sex category. Consequently, there is need for stratification of an area's population into reasonably small homogeneous age, sex and race groups.

It is hypothesized that the supply of net migrants from area A to the rest of the nation or to area A by the rest of the nation is a function of several variables, some of which are measurable and some of which are not. Mathematically we may write:

$$(I) \quad Y = f(X_1, X_2, \ldots X_k, Z_1, Z_2, \ldots Z_n)$$

Where Y represents the supply of net migrants, $X_1, X_2, \ldots X_k$ are k measurable variable and $Z_1, Z_2, \ldots Z_n$ are n non-measurable variables.

This paper is particularly concerned with analyses of internal migration in response to changes in economic activity. Hence, this study investigates the relationship between the rate of unemployment and the rate of internal net migration in a model of the form:

$$(II) \quad Y_{it} = \alpha_i X_t^{\beta_i} Z_t^{\gamma_i} e_{it}$$

where Y represents the rate of internal net migration, X the rate of unemployment, Z the non-observable omnibus variable representing all other variables, and e the residual term, the subscripts i and t denoting area and time interval respectively, and α, β and γ are elasticity parameters. Non-linear iterative least squares estimation procedure is used to estimate the model parameters and the non-observable independent variable Z.

Taking logarithms and minimizing Σe_{it}^2, the system of estimating relations is derived in the usual way. The results are given below. Note that x, y and z are deviations from the mean while corresponding capital letters X, Y and Z represent original values of the variables.

$$\hat{\alpha}_i' = \bar{Y}_i' - \hat{\beta}_i \bar{X}' - \hat{\gamma}_i \bar{Z}'$$

where

$$\bar{Y}_i' = \sum_t Y_{it}'/N_t; \quad \bar{X}' = \sum_t X_t/N_t; \quad \bar{Z}' = \sum_t Z_t'/N_t$$

$$\hat{\beta}_i = \frac{\sum_t z_t'^2 \cdot \sum_t y_{it}' x_t' - \sum_t x_t' z_t' \cdot \sum_t y_{it}' z_t'}{\sum_t x_t'^2 \cdot \sum_t z_t'^2 - (\sum_t x_t' z_t')^2}$$

$$\hat{\gamma}_i = \frac{-\sum_t x_t' z_t' \cdot \sum_t y_{it}' x_t' + \sum_t x_t'^2 \cdot \sum_t y_{it}' z_t'}{\sum_t x_t'^2 \cdot \sum_t z_t'^2 - (\sum_t x_t' z_t')^2}$$

$$\hat{z}_t' = \frac{\sum_i \hat{\gamma}_i y_{it}' - \sum_i \hat{\alpha}_i \hat{\gamma}_i - x_t' \sum_i \hat{\beta}_i \hat{\gamma}_i}{\sum_i \hat{\gamma}_i^2}$$

Empirical Model and Some Data Problems

The model was applied to state time series data covering six decades, 1900-10 through 1950-60, separately to each of the four color-sex categories further subdivided into five age groups. The dependent variable was the rate of internal net migration during the decade. For the measurable independent variable, the "aggregate" value X for each decade was taken as the unweighted arithmetic average of the annual average national unemployment rates. It was recognized that the proper variable to use would be some function of area A unemployment rate and national rate of unemployment or the rate of unemployment in those relevant major occupations in area A where out-migration was taking place and the rate of unemployment in those relevant major occupations in the principal labor markets in the nation into which the net migrant labor was moving. It was unemployment in these particular occupations in which in-migrants engage that was relevant for the purpose. Data considerations, however, precluded the use of such a strictly valid variable. The use of overall national unemployment rate instead would mean that the investigation pertained to the behavioral response of area population to general employment conditions reflecting the phase of the business cycle. [1]

Another important data issue was whether or not theoretical/empirical considerations warranted the imposition of constraints on the minimum and maximum values assumed by X. This aspect of the issue translated into a question of the type: Would the number of net migrants have been materially different if the rate of unemployment in any year during the depression decade was say 10 percent and not 18 percent? It might reasonably be hypothesized that for any given Z, there existed a certain ceiling level at which the adjustment process came to a halt and beyond which higher levels of unemployment would not at all materially affect net migration; and similarly, there existed a certain minimum level beyond which for any given Z, a fall in unemployment rate would not lead to any perceptible increase in the number of net migrants. That is for any given Z, the number of net migrants was influenced by unemployment rate varying within a certain range but it was completely inelastic beyond the end values of this range.

There is empirical evidence in support of these assumptions regarding the existence of a range of variation of X outside which higher or lower values have no further effect on net migration rate. For example, during periods of high unemployment rate, the availability of non-farm jobs to off-farm migrants is sharply reduced. As Schultz (1961, p. 562) observed:

> "The post-war behavior of the economy clearly indicates that the rate of off-farm migration is highly sensitive to changes in unemployment that have characterized these post-war booms and recessions in businesses. Sjaastad's study leaves little room for doubt on this point. Let me put this relationship as follows: when 5, 6 or 7 percent of the labor force is unemployed, the adjustment process under consideration is brought to a halt; on the other hand, when unemployment declines to 3 or 4 percent, off-farm migration becomes large."

Empirical Results

The results of the model using decennial average values for unemployment variable X for all the six decades based on unadjusted annual rates without any upper bound restriction (set (a) Table 1) proved puzzling. Table 2 shows that the signs of the response coefficients β i's pertaining to unemployment rate variable, X, in the case of net out-migration data was negative in nearly 50 percent of the cases. Normally, it should be expected that elasticity of net out-migration with respect to unemployment would be negative; that is, when the rate of unemployment went up, the number of net out-migrants should fall. When $|M_{it}|$ declined, $Y_{it} = 1 + M_{it} / E_{it}$ would rise since M_{it} was negative; and hence there was a positive relationship between Y_{it} and X_t. By identical reasoning one would expect inverse relationship between Y_{it} and X_t in the case of net in-migration. Hence β_i should be positive in net out-migration analysis and negative in net in-migration analysis.

Detailed scrutiny of the empirical results in the case of net in-migration data also showed that a high proportion of β's had positive signs contrary to what one would expect to find on theoretical considerations. Further, the distribution of the signs of β_i's was haphazard and there was no observable pattern of β_i's for individual states. In general, positive and negative β_i's were found for different age groups in each state. [2]

It was considered possible that the unsatisfactory results might be substantially improved if (i) the 1930-40 decade which had a very high value for X was kept out of the regression analyses (set (a) with 1930-40 out) or (ii) an upper limit of say, 7 percent was imposed on individual annual average rates of unemployment and the decennial average calculated accordingly (set (b)). Both these alterations were tried, but the final results did not indicate any substantial improvement and the haphazard distribution of signs of β_i's persisted. The proportion of β_i's,

Table 1. Unemployment Rates, Annual Averages, United States, 1900-1959[a]

Year of Decade	Unemployment Rate (percent) (X)					
	1900-09	1910-19	1920-29	1930-39	1940-49	1950-59
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	5.0	5.9	4.0	8.9	14.6	5.0
1	2.4	6.2	11.9	15.9	9.9	3.0
2	2.7	5.2	7.6	23.9	4.7	2.7
3	2.6	4.4	3.2	24.9	1.9	2.5
4	4.8	8.0	5.5	21.7	1.2	5.0
5	3.1	9.7	4.0	20.1	1.9	4.0
6	0.8	4.8	1.9	17.0	3.9	3.8
7	1.8	4.8	4.1	14.3	3.6	4.0
8	8.5	1.4	4.4	19.0	3.4	6.8
9	5.2	2.3	3.2	17.2	5.5	5.5
Decennial average						
Set (a)[b]	3.7	5.3	5.0	18.3	5.1	4.2[d]
Set (b)[c]	3.5	4.9	4.4	7.0	4.0	4.2[d]

[a]Annual rate of unemployment calculated as number unemployed as percent of civilian labor force.

[b]Set (a): simple arithmetic average of annual values.

[c]Set (b): simple arithmetic average of annual values subject to $X = 7.0$ whenever $X \geq 7.0$.

[d]Annual rates up to and including 1958 are by old definition. Annual rate for 1959 is by new definition.

Sources of Annual Rates:
1900-1954: National Bureau of Economic Research. The Measurement and Behavior of Unemployment, Princeton University Press, 1957; pp. 215-16 (Table 1). Sources as indicated there: 1900-28 present estimates; 1929-39, Monthly Labor Review, July 1948; 1940-54, Bureau of Census. 1954-1958: Bureau of Census, Annual Reports on the Labor Force 1954, 1955, 1956, 1957 and 1958. 1959: Bureau of Labor, Labor Force, December 1959.

having signs contrary to expectation, was still round 40 to 50 percent.

Possible explanations for this unsatisfactory feature of the result might lie in:

(1) Unsatisfactory structural form of the assumed model. The product form of the model might be unsatisfactory insofar as X was concerned. The implicit assumption of constant elasticity might not be appropriate over the range of variation of X covered by the study in set (a). The fact that even considerable narrowing of the range of variation of X by (i) keeping out the 1930-40 decade or (ii) imposing an upper limit constraint on an individual year's unemployment rates at 7 percent did not improve the results might be viewed in support of this possibility. Besides the crucial assumption underlying the iterative procedure that all age groups were confronted with the same X_t and Z_t might not be appropriate and valid.

(2) High degree of correlation between X and Z or between X and some of the other variables contained in Z. Relative wage ratio is one of the major variables contained in Z and X and relative wage ratio might be regarded as being highly correlated. The degree of this correlation between X and Z might vary as between different age groups within a color-sex category in a state, thus resulting in the observed haphazard distribution of signs of β i's among age groups.

(3) Smallness of underlying parameter values. Estimated negative βi's in the case of net out-migration data might be considered as reflecting zero or small positive response coefficient not significantly different from zero.[3] The observed positive βi's in the case of net-in-migration data might be regarded in the same way. This did not, however, appear to be borne out by a detailed analysis of the distribution of signs and of magnitudes of βi's in the case of net out-migration data in Table 2. (Magnitudes of βi's are not shown but the relative picture can be inferred from the affected age groups.)

(4) Unsatisfactory aggregation procedure of using simple arithmetic average to obtain

Table 2 Analysis of Sign of β_i in Two Independent Variable Model, Net Out-migration Data

Region/state	White male age group (i)					White female age group (i)					Nonwhite male age group (i)					Nonwhite female age group (i)				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
New England																				
Maine	+	+	+	+	-	+	-	+	+	-										
New Hampshire		+	+	+	+	+	-	-		-										
Vermont	-	+	+	+	-	+	-	+	+	-										
Massachusetts	+		+	-	+	+		+	-	+										
Rhode Island			-	+	-															
Middle Atlantic																				
Pennsylvania	+	-	-	+	+	+	-	-	+	-										
East North Central																				
Indiana		-	+		-		+	+	+	-										
Illinois	-			-	+	-		-	-	+										
Wisconsin	+	+	-			+	+	+												
West North Central																				
Minnesota	-	+	-	-		-	-	-	+											
Iowa	-	+	+	-		+	-	-	+	+										
Missouri	+	-	-	+	+	-	-	-	+	+										
North Dakota	+	+	-	-	+	+	-	+	+	+										
South Dakota	+	+	+	-	+	+	-	+	-	+										
Nebraska	+	+	-	+		+	-	+	+	+										
Kansas	-	-	+	+		-	-	+	+	-										
South Atlantic & D. C.																				
Maryland											+			-	-	+			+	-
Dist. of Columbia	-		+	-																
Virginia							-	+	+	+	-	-	+	+	-	+	-	+	+	-
West Virginia	+	+	-	+		+	+	+	-	-	-		-	-	+	-	-	-	-	+
North Carolina	-	-	+	+	+	+	+	+	-	+	-	+	+		-	+	+	+	+	-
South Carolina		-	+	-	-	+	+	+	-	+	+	+	-	+	-	+	+	-	-	+
Georgia	-	+	+	-	-	-	-	+	+	+	+	+	+	+	-	+	+	+	-	+
East South Central																				
Kentucky	-	-	-	+	+	-	-	-	+	+	+	-	+	+	-	+	-	+	+	-
Tennessee	+	-	-	-	+	+	+	+	+	+	+	-	+		+	+	+	-	+	-
Alabama	+	+	-	-	+	+	+	+	-	-	+	+	+	-	+	+	+	+	-	
Mississippi	-	-	+	+		-	-	+	+	+	-	+	-	-	-	-	+	+	-	-
West South Central																				
Arkansas	+	+	-	+	+	+	-	-	+	+	+	+	+		-	-	+	+	+	-
Louisiana		-		+	+						-	+	-	+		-	+	-	+	
Oklahoma	+	+	+	-		+	+	+	-		+	+	+			+	+	+		
Texas											-	+	+				+	-	-	
Mountain																				
Montana	-	+	+	-	-	+	+	-	-	+										
Idaho	-	+	-	+		+	+	+	-	-										
Wyoming	+		-	+	-	+	+		+	-										
New Mexico						-	+		-	-										
Utah	+	+	-	+																
Total +	15	17	14	18	13	20	12	19	17	16	8	9	9	5	3	9	10	8	7	3
-	12	10	15	12	9	8	16	9	11	11	6	3	4	4	8	4	3	5	6	7
Grand Total	27	27	29	30	22	28	28	28	28	27	14	12	13	9	11	13	13	13	13	10

decennial value for X from the individual year's data. The unsatisfactory character might lie either in the type of average used (a geometric average might be regarded as more appropriate in a multiplicative model) or in the use of equal weights. Since individual year's X's varied greatly over each decade, the problem of weighting, when ignored, could cause serious distortions.

(5) Unsatisfactory nature of the national rate of unemployment as a valid proxy explanatory variable reflecting demand relevant for individual states.

Further Search for a Plausible Explanation

The empirical evidence in support of the indeterminate sign of the coefficient of elasticity of Y with respect to X was so overwhelming that it was necessary to look for suitable plausible hypothesis or hypotheses to explain the result.

Reverting to the initial relationship between Y, X and Z, we have:

(III) $Y = \alpha X^{\beta} Z^{\gamma}$

If X and Z are assumed to be independent, elasticity of Y w. r. t. X, $\epsilon^{(x)}$ is equal to β. If, however, X and Z were highly correlated, problems of multicollinearity might arise. Let Z be regarded as a function of X and ζ where X and ζ are independent. Let $Z = f(x, \zeta) = X^{\delta} \zeta^{\lambda}$. (III) may now be written as

(IV) $Y = \alpha X^{\beta + \delta\gamma} \zeta^{\lambda\gamma}$ Hence:

(V) elasticity of Y with respect to X, $\epsilon^{(x)} = \beta + \delta\gamma$

(VI) and elasticity of Y with respect to ζ, $\epsilon^{(\zeta)} = \lambda\gamma$

In this situation, the observed value of the response coefficient pertaining to X is really the value of $\epsilon^{(x)} = \beta + \delta\gamma$ and the observed value of the response coefficient pertaining to Z is really the value of $\epsilon^{(\zeta)} = \lambda\gamma$.

Hypothesis Regarding the Relationship Between Z and X

Z, the omnibus nonobservable variable is the overall representative of all possible explanatory variables affecting Y excepting X.

$$Y = \alpha X^{\beta} Z_1^{\gamma_1} \ldots Z_k^{\gamma_k} \text{ and}$$

$$Z^{\gamma} = Z_1^{\gamma_1} Z_2^{\gamma_2} \ldots Z_k^{\gamma_k}$$

Some of the Z_k's may have a high degree of correlation with X. In the previous section, what has essentially been done is to separate out all those Z_k's which are independent of X from those that are correlated with X and to replace the latter group by $X^{\delta} = Z_1^{\gamma_1} \ldots Z_j^{\gamma_j}$ and writing the remaining group by $\zeta^{\lambda} = Z_{j+1}^{\gamma_{j+1}} \ldots Z_k^{\gamma_k}$. It is important to note that nothing has yet been assumed for δ; some of the Z_j variables correlated with X may be inversely correlated and hence some of the γ_j's may be negative. Consequently δ which is a function of γ_j's may be positive or negative.

Consider the following hypothesis: Over the business cycle, the amplitude of fluctuations in the index of relative opportunity in net in-migration areas is less than that in areas of net out-migration.

It may be argued that as the rate of unemployment in the national economy increases, downward pressure on wages and other opportunity factors will be felt in both types of areas. In a net in-migration area, however, the burden of adjustment will partly be borne by potential in-migrants and will be felt in reduced net in-migration. Potential net in-migrants thus serve as an initial safety valve against the downward pressures on wages and other opportunity factors in net in-migration areas. In a net out-migration area, on the other hand, the downward pressure on wages, etc., caused by a general decline in economic activity is additionally reinforced by reduced out-migration. Similarly, in an upswing, the upward pressures on wages, etc., in net in-migration areas are partly neutralized by increased net in-migration. In a net out-migration area, on the other hand, the upward pressures on wages, etc. as a result of general economic expansion are reinforced by increased pace of out-migration.

On the above reasoning, the hypothesis states that during the upswing of the business cycle there is less tendency for the index of relative opportunity to rise in net in-migration areas than in net out-migration areas. Since the 'rest of the nation' in relation to a net in-migration area A will include net out-migration areas also, we would assume that wages and opportunity factors in area A rise less than the rest of the nation during the upswing of the business cycle. Z, the index of relative opportunity, will thus tend to fall. Similarly, for area A, a net in-migration area, wages and other opportunity factors will tend to fall less than in the rest of the nation in the declining phase of the business cycle. This means that the index of relative opportunity will tend to rise. Conversely, for a net out-migration area B, during the upswing of the business cycle, the upward pressure on wages and other opportunity factors will be more than that in the rest of the nation; hence, the index of relative opportunity would tend to rise.

In the declining phase of the business cycle, the downward pressure on wages, etc., in these areas will be more than in the rest of the nation, so that Z would tend to fall. The above hypothesis is, therefore, equivalent to $\delta<0$ for net out-migration and $\delta>0$ for net in-migration areas.

Let us use subscript (1) to denote net out-migration analyses and subscript (2) to denote net in-migration analyses. Let $\varepsilon_{(1)}^{(x)}$ and $\varepsilon_{(1)}^{(\zeta)}$ denote elasticity of Y with respect to X and ζ, respectively, for net out-migration areas; similarly let $\varepsilon_{(2)}^{(x)}$ and $\varepsilon_{(2)}^{(\zeta)}$ refer to net in-migration areas. We have

$$\varepsilon_{(1)}^{(x)} = \beta_{(1)} + \gamma_{(1)}\,\zeta_{(1)};\quad \varepsilon_{(1)}^{(\zeta)} = \gamma_{(1)}\,\lambda_{(1)}$$

$$\varepsilon_{(2)}^{(x)} = \beta_{(2)} + \gamma_{(2)}\,\zeta_{(2)};\quad \varepsilon_{(2)}^{(\zeta)} = \gamma_{(2)}\,\lambda_{(2)}$$

For net out-migration areas, $\beta_{(1)}$ is expected to be positive; but $\delta_{(1)}$ is expected to be negative on the basis of the hypothesis advanced in the earlier section. The sign of $\varepsilon_{(1)}^{(x)}$ will, therefore, be indeterminate only if $\gamma_{(1)}$ is positive. Similarly, for net in-migration areas, $\beta_{(2)}$ is expected to be negative; but $\delta_{(2)}$ is expected to be positive on the basis of the hypothesis of the earlier section. The sign of $\varepsilon_{(2)}^{(x)}$ will, therefore, be indeterminate only if $\gamma_{(2)}$ is positive. Thus, for both types of areas, a hypothesis that leads to $\gamma>0$ together with the hypothesis of the earlier section will serve to explain satisfactorily the observed haphazard distribution of ε's, i.e., of the coefficient associated with X in the empirical results.

γ is the power of Z term in the basic relationship $Y = \alpha \times {}^{\beta}Z^{\gamma}$. $\gamma>0$ implies direct relationship between Y and Z. One important variable covered in Z is the relative wage ratio and we may reasonably regard this as the dominant variable included in Z. The number of people who live in an area, divided by appropriate exposed to risk, Y, and the relative wage ratio, Z, are, by the following reasoning, directly related in net out-migration and net in-migration areas.

Thus, there exists a reasonably valid basis for the observed haphazard distribution of signs of the response coefficients associated with X. The hypothesis of the previous section together with expected positive sign of γ yields the desired results. $\varepsilon_{(1)}^{(x)} = \beta_{(1)} + \delta_{(1)}\gamma_{(1)}$, $\beta_{(1)}>0$, $\gamma_{(1)}>0$ and $\delta_{(1)}<0$; hence the sign of $\varepsilon_{(1)}^{(x)}$ is indeterminate. Similarly $\varepsilon_{(2)}^{(x)} = \beta_{(2)} + \delta_{(2)}\gamma_{(2)}$, $\beta_{(2)}>0\gamma_{(2)}>0$ and $\delta_{(2)}>0$; hence the sign of $\varepsilon_{(2)}^{(x)}$ is indeterminate.

Elasticity of Y with Respect to ζ

The empirical results showed that a very high proportion of coefficients connected with ζ carried a positive sign. $\varepsilon_{(j)}^{(\zeta)} = \gamma_{(j)}\,\lambda_{(j)}$ (j = 1, 2) and $\gamma_{(j)}>0$ (j = 1, 2). Hence $\lambda_{(1)}, \lambda_{(2)}>0$. Consequently, it must be assumed that Z and ζ are positively associated both in the case of net out-migration and net in-migration areas. ζ is the aggregate index of the net effect of all variables affecting wage ratio other than employment rate X, after the effect of X has been separated out of all those variables.

Unanswered Questions

In the above discussion, it has been implicitly assumed that the iterative procedure applied to Y, X data yields estimates of ζ, where ζ and X are independent. Little is known of the properties of the parameter estimates yielded by the two variable models or about the character of the nonobservable variable whose estimates are thrown up by the iterative process. In the nature of things, no empirical basis can exist for proving the character of ζ and its independence or otherwise of X.

Footnotes

[1] Segal (1962) faced a similar problem in investigating the influence of the strength of the demand for labor on occupational wage differentials. In the absence of reliable data pertaining to unemployment in individual areas, he also used national rates.

Jerome (1926, p. 54) observes: The cycle of employment is the aspect of the business cycle which is of direct meaning to the immigrant. It is the most tangible measure of the conditions affecting his economic welfare; and hence it affords the obvious and logical basis for appraising the influence upon migration of fluctuations in economic opportunities and the celerity with which immigration and emigration currents respond to such changes.

Jerome (1926, p. 121) further observes that: Inasmuch as good employment conditions would presumably encourage the prospective immigrant, ... we may reasonably assume, ... that business conditions are in fact a dominating determinant of cyclical fluctuations in immigration.

[2] Sjaastad (1961 p. 50) in his analysis of income and net migration in the United States also ran into a comparable situation. He found that "the unemployment coefficient, although erratic in

sign, is negative whenever significant, implying paradoxically that higher level of unemployment attracts larger shares of migrants; however the causation is probably the other way around, with the larger shares of migrants contributing to unemployment."

[3] Johnston's (1963) analysis also yielded some negative βi's contrary to expectation, but he regarded them as essentially nonsignificant.

References

Agricultural Policy Institute. 1961. "The farmer and migration in the United States." API Series No. 3, School of Agriculture, North Carolina State College, Raleigh.

Bishop, C. E., 1961. "Economic Aspects of Changes in Farm Labor Force," pp. 36-49. In Labor Mobility and Population in Agriculture. Iowa State University Press, Ames.

Bunting, R. L., 1962. "Labor Mobility and Wage Improvement," pp. 208-219. Conference on Human Resources in the Urban Economy. Resources for the Future, Inc., Washington, D. C.

Diehl, W. D., 1964. Farm-nonfarm Migration in the Southeast: A Costs-returns Analysis. Unpublished Ph. D. thesis, Department of Agricultural Economics, North Carolina State of the University of North Carolina at Raleigh. University Microfilms, Ann Arbor, Michigan.

Jerome, H., 1926. "Migration and Business Cycles." National Bureau of Economic Research, Inc., New York.

Johnston, W. E. and Tolley, G. S., 1968. "The Supply of Farm Operators." Econometrica, Vol. 36, No. 2 (April 1968) pp. 365-382.

Kripalani, G. K., 1969. "Internal Net Migration Response Differentials for the United States by Nonlinear Least Squares Estimation Procedure," Proceedings of the American Statistical Association, Social Statistics Section, Washington, D. C.

Kripalani, G. K., 1970. "Race-Sex Discrimination in Internal Net Migration Proceedings of the American Statistical Association," Social Statistics Section, Washington, D. C.

Kripalani, G. K., 1972. "Race-Sex Discrimination Indices for Non-metropolitan State Economic Areas," Proceedings of the American Statistical Association, Social Statistics Section, Washington, D. C.

Schultz, T. W., 1961. "A Policy to Redistribute Losses from Economic Progress." J. Farm Economics, 43 (3): pp. 554-565.

Segal, M., 1962, "Occupational Wage Differentials in Major Cities During the 1950's, pp. 195-207. Conference on Human Resources in the Urban Economy. Resources for the Future, Inc., Washington, D. C.

Sjaastad, L. A., 1961a. Income and Migration in the United States, Unpublished Ph. D. thesis, Department of Economics, University of Chicago. University Microfilms, Ann Arbor, Michigan.

Sjaastad, L.A., 1961b. "Occupational Structure and Migration Patterns," pp. 8-27. In Labor Mobility and Population in Agriculture. Iowa State University Press, Ames.

MORTALITY DECLINE IN INDIA, 1951 - 1961: DEVELOPMENT VS PUBLIC HEALTH PROGRAM HYPOTHESIS

P. Krishnan
Population Research Laboratory
Department of Sociology
University of Alberta
Edmonton

The decline in mortality in the developing nations is usually attributed to public health programs and imported western medical technology. Thus Kingsley Davis speaks of the "amazing decline" of mortality in underdeveloped countries. In the case of Ceylon (Sri Lanka), an oft quoted example for this "amazing decline", this has been questioned by Fredericksen. According to Fredericksen, it is economic development that is responsible for, or associated with, the decline in mortality in Ceylon. The case of India is examined here employing the path analysis technique. The analysis of cross-section data indicates that the mortality decline in India might have been due to, broadly speaking, development and not primarily a consequence of public health programs.

1. INTRODUCTION

It is an almost accepted fact in the demographic literature that the decline in mortality in the developing countries is due to the importation of western medical technology and public health programs. Research done in this area reveals that mortality fell most rapidly in western countries in the late nineteenth and twentieth centuries, and the decline was largely the resultant of medical progress achieved in disease control. But the long secular decline in mortality, before this period, largely emanated from economic improvements such as increasing agricultural efficiency, the introduction of superior varieties of crops and live stock permitting better diets, and improvements in transporation eliminating famines due to local food shortages (Wrong, 1967: 41). Demographers (Sociologists) are of the opinion that the mortality decline(s) realized in developing countries are independent of an overall economic and social modernization. Thus Davis (1956) speaks of the amazing decline in mortality in non-western societies without their undergoing thorough transformation of social and economic structures. The case of Ceylon (Sri Lanka) cited by him has been examined by Fredericksen (1960) who seems to doubt the contribution of malaria control campaign in reducing the mortality level. Fredericksen's (1961) further analysis shows that economic development by increasing per capita food consumption was an important cause for the mortality decline in that country. This study on Sri Lanka has to be viewed as an eye-opener for demographers working on developing countries, for these nations are, inter alia, undergoing economic and social modernization. Also it seems to add some precepts to an area which is conspicuous by paucity of adequate theory. Stolnitz (1955) who has drawn generalizations on international mortality trends, echoes this remarkably.

> Increasing life chances are almost always explained by reference to two broad categories of causes; rising levels of living on the one hand (income, nutrition, housing, literacy), and on the other hand technological advances (medical science, public health, sanitation). The usual approach has been to regard these sets of factors as more or less coordinate, with little attempt to assess their relative importance. At the same time there has been considerable emphasis on their interdependence, a common observation being that the development and the application of disease-control techniques would have been very different in the absence of widespread social change.
>
> Both of these views, which evolved largely on the basis of western mortality experience, have also been traditional explanations of the contrasting patterns found in other parts of the world. Only recently has their adequacy been seriously questioned, mainly as a result of developments in Latin America - Africa - Asia. The introduction of new disease-control techniques in this region, usually unaccompanied by shifts in socio-economic conditions, has led to drastic mortality declines in the last few years. It is worth noting, therefore, that a similar process may have been operative in the acceleration of western survivorship a good deal earlier.

The objective of this paper is to look at the mortality situation in India for the period 1951-61 and conclude on its possible determinants Based on an analysis of mortality of the different states in India, Kohli (1971) is of the opinion that mortality decline in India has been mainly due to public health programs. The present analysis is an attempt to examine the validity of the Kohli conclusion.

2. METHODOLOGY

The study proposed in this paper is a cross sectional rather than a temporal analysis. There are several reasons for choosing a cross-section probe. A temporal analysis, will (may) not separate out the major agents of mortality decline as many events are taking place simultaneously. The different states in India show different levels of mortality, economic, and social development. Assuming that the change in mortality from a high to a low level is the resultant of the different forces of social and economic modernization, it is possible to separate out the influence of each

of the major factors on the decline of mortality. The analysis is done with the help of single equations. Since the objective of this study is to elicit the causes of mortality decline, instead of the conventional regression analysis, path analysis framework (Land, 1969; Wright, 1960), is employed.

We propose two competing path models from substantive considerations and test them for adequacy.

Model I.

Many social scientists tend to confine development to the economic dimension only; and all others being effectuated by it. If we adhere to this view on developmental process, we have the following path model for mortality (decline) in a country.

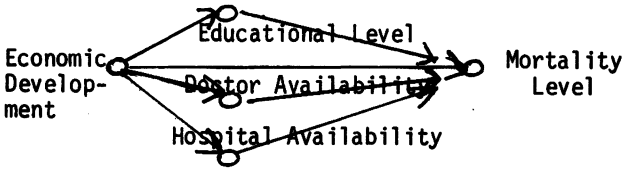


Fig. 1: ECONOMIC DEVELOPMENT-ORIENTED MODEL

In this model, economic development directly introduces changes in mortality level through a high standard of living. It also indirectly influences mortality through factors such as education, hospital services, etc. We use the following indicators.

Variables	Indicator	Label
Economic Development	State Income Per Capita	Z_5
Education	State Literacy Rate	Z_4
Doctor Availability	State Doctor-Population Ratio (Number of people served by one doctor)	Z_3
Bed Availability	State Bed-Population Ratio (Number of people served by one hospital bed)	Z_2
Mortality	Crude Death Rate	Z_1

Per-capita food intake can be included here to make the model more comprehensive. The path diagram with these variables is available in Figure 2.

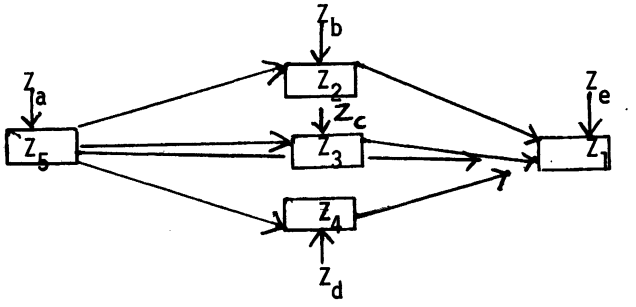


Fig. 2: PATH DIAGRAM FOR MODEL 1

Legend:

Z_5 - Income	Z_a	Error terms
Z_4 - Literacy	Z_b	
Z_3 - Doctor/Population Ratio	Z_c	
Z_2 - Bed/Population Ratio	Z_d	
Z_1 - Death Rate	Z_e	

Z's are standardized variables. Under the usual assumptions underlying path-regression analysis, the following equations can be derived.

$$Z_5 = p_{1a} Z_a$$

$$Z_4 = p_{45} Z_5 + p_{4d} Z_d$$

$$Z_3 = p_{35} Z_5 + p_{3c} Z_c$$

$$Z_2 = p_{25} Z_5 + p_{2b} Z_b$$

$$Z_1 = p_{15} Z_5 + p_{14} Z_4 + p_{13} Z_3 + p_{12} Z_2 + p_{1e} Z_e$$

This yields the normal equations I & II

$$r_{25} = p_{25}$$

$$r_{35} = p_{35} \qquad \text{I}$$

$$r_{45} = p_{45}$$

$$r_{15} = p_{15} + p_{14} r_{45} + p_{13} r_{35} + p_{12} r_{25}$$

$$r_{14} = p_{15} r_{45} + p_{14} + p_{13} r_{34} + p_{12} r_{24} \qquad \text{II}$$

$$r_{13} = p_{15} r_{53} + p_{14} r_{43} + p_{13} + p_{12} r_{23}$$

$$r_{12} = p_{15} r_{52} + p_{14} r_{42} + p_{13} r_{32} + p_{12}$$

Furthermore, we have,

$$r_{23} = p_{25}\, p_{35} = r_{25}\, r_{35}$$

$$r_{24} = p_{25}\, p_{45} = r_{25}\, r_{45} \qquad \text{III}$$

From set (II), we can solve for the path coefficients. The equations in set (III) aid in examining the empirical adequacy of the model.

Model II.

Non-economists view development from a broader perspective, economic development being only one of the essential ingredients. Changes in social institutions and structures, political stability etc. have, a great role to play in the modernization of the Third World.

Achievements registered in educational sector, public health programs etc. will be considered as elements of the developmental activity and as separate inputs into the system. Growth in income is the usual indicator of economic development. For our analysis, political changes need not be considered. The political stability factor is assumed to remain the same throughout the period under consideration.

A suitable path model in this case is obtained considering all the developmental activities as exogenous variables operating on the mortality factor. The path diagram is shown in Figure 3.

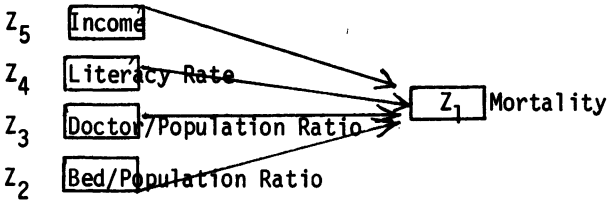


Fig. 3: THE GENERAL DEVELOPMENT PATH MODEL

In this model, we do not analyse the interrelationships of the exogenous variables.

The above model yields the conventional multiple regression situation except that it is causally interpretable. The following normal equations determine the path coefficients.

$$r_{12} = p_{12} + p_{13}\, r_{32} + p_{14}\, r_{42} + p_{15}\, r_{52}$$

$$r_{13} = p_{12}\, r_{23} + p_{13} + p_{14}\, r_{43} + p_{15}\, r_{53}$$

$$r_{14} = p_{12}\, r_{24} + p_{13}\, r_{34} + p_{14} + p_{15}\, r_{54} \qquad \text{IV}$$

$$r_{15} = p_{12}\, r_{25} + p_{13}\, r_{35} + p_{14}\, r_{45} + p_{15}$$

The set of equations (IV) is identical with the set (II). The distinction between models I and II arises from the substantive bases of the two leading to a set of conditions given in III for model I only. The constraints as given by III determine whether model I is consistent with the data or not.

3. DATA SOURCES

The data used in this analysis come from a variety of sources. Since there is under-registration of vital events in India, even death rates for the states have been estimated with the help of census age distributions. Agarwala (1967) presents several estimates of crude death rate and expectations of life at birth for the different states in India. The quasi-stable estimates of death rates have been used for the exercise attempted here. Literacy rates are again from the census returns (Agarwala, 1967: 51). Doctor-population and bed-population ratios have been computed from the data provided in the vital statistics of India 1961 (India, 1963). These data (see Table 4) were available for only eleven states.

It is worthwhile to know some of the limitations of the data and the methodology employed here. Clearly the kind of data with which the analysis is performed, is macro and all the limitations associated with official data can be stated here. No time lag is allowed for most of the independent variables as far as their effect on the dependent variable is concerned. This is because such types of information cannot be secured easily. Furthermore, in a cross-section analysis where regions are the units of analysis, the question of ecological correlations arises. This need not concern us as our aim is only to seek the relevant determinants of differential mortality with regard to such units of analysis. The sample size is small. It would have been better to take the districts of India as the units of analysis. Even with states as the units, it was difficult to gather data for all the seventeen states in India. So one has to stay with the problem of small sample size. Inferences drawn have to be taken with some caution.

4. DATA ANALYSIS

To start with, a non-causal multiple regression analysis was performed to select the relevant independent variables. Indicators of economic development (per capita state income), health programs (bed-population, doctor-population, hospital-population, dispensary population ratios, per capita state expenditure on medical and health services), social development (per cent population urban) and social change (literary rate) were employed as the independent variables. The performances of these variables were approximately the same when the dependent variable was the crude death rate or the expectation of life at birth. In view of

its simplicity, the analysis is being restricted to crude death rate as the dependent variable.

The step-wise regression yielded the following results.

Variable Added	Percent Variation Explained	Increase
Literacy Rate	47.03	47.03
Bed-Population Ratio	58.65	11.63
Doctor-Population Ratio	65.35	6.70
Per Capita Income	70.85	5.50
Hospital-Population Ratio	74.49	3.64
Dispensary-Population Ratio	78.58	4.09
Urbanization	84.41	5.83
State Per Capita Expenditure in Medical and Health Area	86.34	1.93

If 5 per cent level of significance is considered, the regression coefficients of the first three independent variables (literacy rate, bed-population and doctor-population ratios) are significant. All others (including in particular, per capita expenditure on medical and health services) do not add any significant regression components at all. Since there is considerable interest on income as such, it was decided to keep the first four variables for the purposes of this study. About 70.9 per cent of the variation in the dependent variable is accounted for by these four variables.

The four variables picked are significant ones from a developmental perspective. Bed-population and doctor-population ratios are indicators of the progress of the medical and health programs. Income per capita is an indicator of economic growth and literacy rate is an indicator of the social awareness of the population and also an agent of social change. The roles of these independent variables are sociologically interpretable.

Adequacy of Model I.

As stated earlier, both the models are based on substantive considerations. The adequacy of the models depends on how they fit with the data.

a. Model I. The equations comprising set III and II are the constraints which have to be satisfied by the empirical fit. We have (see Table 1), the following.

Since the expected and the observed correlations differ by a wide margin, the fit of the model under consideration is not satisfactory.

b. Model II. In this model, economic growth and other inputs are considered as separate forces determining a mortality situation. The observed and the expected correlations are presented below.

The model fits with the data well and hence can be preferred to Model I to explain the changing mortality situation in India.

4. DISCUSSION

Since Model I is not an adequate representation of the forces at work leading to differential regional mortality in India, the economists' contention that everything follows economic growth is not tenable. It also suggests that per capita state income (GDP) is a poor indicator of the development achieved by a province (state). Model II characterizes the mechanism behind the changing mortality situation in India. The indicators used here are those of economic development (per capita state income (GDP), public health programs (doctor/bed-population ratios) and the social awareness and change factor (literacy rate).

In path analysis, the role of these factors can be looked at from a direct and an indirect perspective. The direct effect (indicated by the path coefficients) measures the influence produced by a factor on the dependent variable by itself, while the indirect effect is the influence through other factors. Finney (1973) introduces a causal connotation to indirect effect and defines 'a causal indirect effect.' In view of the kind of setup we have in Model II, we cannot develop "indirect causal effect" estimates. The direct causal and the residual effects of these developmental factors on mortality are presented in Table 3. The indirect effects, in this situation, are not causally interpretable.

It is clear from the path coefficients given in Table 3 that the total direct effect of public health program is -.070 while those of literacy and income are much larger in magnitude. The results are in the expected direction. Even when treated separately, the roles of literacy and income are not negligible as compared to the doctor-population factor. It can be noted that the indirect effect of literacy on

mortality is almost as large as the direct effect of doctor/population ratio. Literacy (education) has a two-fold role to play. On the preventive side, literacy helps to alleviate mortality by promoting social hygiene and on the control side, the awareness generated is capitalized in making the best use of the medical facilities provided by the health program. Similar reasonings can be drawn for the direct and the indirect effects of income. A higher average per capita consumption of food leading to a more intake of cereals, may be a consequence of higher per capita GDP. This produces a reduction in mortality. Indirect effects of income are evinced through spending capacity on medical expenses.

5. CONCLUSION

From the analysis presented here, it is clear that the state differentials in mortality and hence the decline in mortality in India are a consequence of not simply public health programs only. Economic development and social change factors also had a significant contribution in reducing mortality in the period 1951-1961. The inevitable conclusion is that the public-health program hypothesis regarding mortality decline with respect to the developing nations is not tenable in some cases. Generalizations in demography regarding mortality decline need revision in the light of such experiences.

REFERENCES

Agarwala, S.N. (1967), Population - India: The Land and the People, New Delhi: National Book Trust.

Davis, K. (1956), "The amazing decline of mortality in underdeveloped areas", American Economic Review, 46, 305-318.

Finney, J.M. (1973), "Indirect effects in path analysis", Sociological Methods and Research, 1: 175-186.

Frederiksen, H. (1960), "Malaria control and population pressure in Ceylon", Public Health Reports, 75(10) - Reprinted in Readings on Population (ed.: David M. Heer), Englewood Cliffs: Prentice-Hall, 69-73.

Frederiksen, H. (1961), "Determinants and consequences of mortality trends in Ceylon", Public Health Reports, 76(80) - Reprinted in Readings on Population, 74-80.

India, Government of (1963), Vital Statistics of India 1961, New Delhi: Ministry of Home Affairs.

Kohli, K.L. (1971), "Spatial Variations of Mortality in India, 1951-61", Dissertation Abstracts, 32, 4733A.

Land, K. (1969), "Principles of path analysis", Sociological Methodology 1969 (eds. E.F. Borgatta and G.W. Bohrnstedt), San Francisco: Jossey Bass, 3-37.

Stolnitz, G.J. (1955), "A Century of international mortality trends: I", Population Studies, 9, 26-55.

Wright, S. (1960), "Path coefficients and path regression: alternative or complementary concepts?", Biometrics, 16, 189-202.

Wrong, D.H. (1967), Population and Society, New York: Random House, Inc.

Table 1

OBSERVED AND EXPECTED CORRELATIONS FOR MODEL I

Corr. Coef.	Observed (1)	Expected (2)	Difference Col. 1-Col. 2
r_{23}	.156	.300	-.144
r_{24}	.262	-.147	+.409
r_{34}	-.552	-.164	-.388
r_{15}	-.405	-.893	+.488

Source: Statistical Analysis

Table 2

OBSERVED AND EXPECTED CORRELATIONS FOR MODEL II

Corr. Coef.	Observed	Expected	Difference
r_{12}	-.289	-.289	0
r_{13}	.663	.662	.001
r_{14}	-.686	-.685	-.001
r_{15}	-.405	-.405	0

Source: Statistical Analysis

Table 3

DIRECT AND INDIRECT EFFECTS ON MORTALITY, INDIA 1951-61

Development Factor	Direct Effect (Path Coef) On Mortality	Indirect Effect (Path Coef) On Mortality	Indirect Effect Through
Economic Development Factor			
Income	-.349	-.056	Doctor/Pop. .244 Bed/Pop. -.231 Literacy -.069
Public Health Program Factor			
Doctor/Population	-.469	.180	Income .182 Bed/Pop. .062 Literacy -.064
Bed/Population	.399	.264	Income .202 Doc/Pop. -.073 Literacy .134
Social Change Factor			
Literacy	-.243	-.443	Doc/Pop. -.123 Income -.099

Table 4

DATA EMPLOYED FOR ANALYSIS

	Deathrate 1951-1961	Per Cent Literate 1961	Population In Thousands Served By One Bed 1961	Population* In 1000's Served By One Doctor	Per Capita State Income (R s)
Assam	26.9	27.4	3.40	3.81	330.96
Andhra Pradesh	25.2	21.2	1.79	11.14	296.13
Kerala	16.1	46.8	1.40	20.92	305.03
Madhya Pradesh	23.2	17.1	3.25	13.66	274.73
Madras (Tamilnadu)	22.5	31.4	1.40	2.90	359.95
Maharastra	19.8	29.8	1.43	3.52	398.81
Mysore	22.2	25.4	1.53	6.73	257.08
Orissa	22.9	21.7	3.19	12.25	237.05
Punjab	18.9	24.2	1.53	3.73	432.19
Uttar Pradesh	24.9	17.6	2.80	9.89	272.75
West Bengal	20.5	29.3	1.20	7.20	461.09

* Data on registered medical practioners refers to 1959 and the population to 1961. The rate for Maharastra refers to old Bombay State which included the present states of Maharastra and Gujarat.

Source: 1 Statistical Abstract of Indian Union 1960, Central Statistical Organization, New Delhi

2 Health Statistics of India 1961 and 1962, Ministry of Home Affairs, New Delhi

TESTIBILITY OF HYPOTHESES THROUGH INTERVIEW TAPING AS MEANS OF SOCIAL SURVEY EVALUATION

Karol J. Krótki, University of Alberta

A completed questionnaire from a social survey is both an incomplete and overperfect record of what transpired between the respondent(s) and the interviewer. The recorded answers abstract from the interview dynamics and simplify the complexity of answers into categories acceptable to the supervisors of the interviewer, even in the case of open ended questions, and certainly so in the case of pre-coded answers. The written entries throw no light on the process through which the respondent arrived at his answers, except when the interviewer offers marginal comments, or supplementary questions ask for the recording of any documents that might have been consulted (e.g., tax statements in the case of income). Completed questionnaires give no information on the "cost" of a particular question in terms of time and strain on the goodwill between the two parties. There are no means on the questionnaire to indicate how far the interviewer was forcing the conversation and answers to his questions into moulds strange, ill understood by the respondents and away from what the respondent may have wanted to talk about.

Discipling the respondents' answers does not end with whatever forcing the interviewer engaged in. It is continued through the processes of mechanical and manual editing of data often in line with plausible and preconceived values for the different parameters. The existence of mathematical and theoretical models lends credence and support to such activities and culminates in further adjustment to data, both at the micro and at the macro level. However, this paper is not concerned with these latter influences.

In an earlier paper techniques of analyzing taped interviews were assessed (Krótki, 1973). Briefly, with a survey of any size (that is excluding short journalistic or psychological interviews with individuals), transcription seems to be the quickest and least expensive step leading to categorizations, coding systems, and aggregations.[1] This will be unwelcomed news to those who experienced the painful and bothersome procedures of transcription and were looking forward to some alleviation from their travails. In this paper we move a step further and ask what hypotheses can be tested through tape recording on the assumption that the taped interviews can be suitably processed. Such processing would include transcribing, categorizing relevant features of the taped interview, coding, quantifying wherever appropriate, and aggregating for suitable comparisons with other data, including questionnaires from interviews not tape recorded. There are dozens of little tricks of trade, ignorance of which makes taped interviews that much less valuable, all of which are also outside this paper.[2]

Verbatim recording of interviews and the subsequent use of the recordings can have a number of purposes and can be done in a variety of ways. Transcription on location during the interview even by skilled stenographers loses one third to one half of the interview due to conscious and unconscious selection of what to write (Bucher et al., 1965a; 1965b). When note taking is considered more obtrusive than tape recording during clinical, counselling, and psychological interviews the value of the tape lies in the possibility of almost limitless replay and unhurried consideration of each detail of the conversation (Brody et al., 1951; Eitzen, 1952). Tape recordings are being used to record hot news and immediate reactions. Such recordings, however, are seldom done on a large scale with structured questionnaires. In this paper we are interested not only in repetitive interviews leading to structured records, but also in the limited aim of evaluation, i.e., not in recording to produce substantive data. Audio-visual recordings are used when it is felt that some of the dimensions of the interview dynamics escape audio tapes (Kantner & Zelnik, 1969).

In the process of evaluation some data correcting procedures may be refined, but we resist the endless temptations to fork out in the pursuit of anthro-cultural and socio-psychological curiosities that arise in almost every minute of many an interview. Our purpose is to report on the testibility of objective and quantified hypotheses through tape recorded interviews. A number of hypotheses in six taxonomic groups have been chosen and their potentialities or actual experiences, where already available, are reported upon.[3]

The first question that arises with tape recordings is how far the process distorts the interview. The distortion can be two-fold: relatively to other non-taped interviews and relatively to actual "facts" as otherwise reportable. The overwhelming opinion of participants seems to be that there is little distortion (brief summary in Krotki, 1973). These considerations pose the first three hypotheses, all three methodological (M) and tape (T) related. They are followed by other methodological (M) hypotheses that concern survey (S) questions as such arising without tapes.

(MT 1) Tape recording does not distort the content of the interview relatively to interviews not tape recorded, though as suggested in hypotheses (MT 2) the effect of recording is not entirely neutral.

A rigorously conducted experiment involving a priori interpenetrated samples, or less certainly a posteriori matched groups, should provide definite answers anthropological culture by culture.[4] Instead, we have in literature less rigorous enquiries ending impressions.

This writer can add to them his own surprise how unobtrusive is the obtrusiveness of the microphone. Protruding right in the middle of the conversing group it seemed to him to have been quite neutral in the many Moroccan homes, where the taping operation was observed during the population census of 1971. To respond objectively and definitely to hypothesis (MT 1) it would be extravagent to study each substantive characteristic. It might be enough to obtain some objective indicator such as the number of interviews conducted during the same unit of time by the same interviewer in comparable circumstances when taped and when not taped. Another possible indicator could be the proportions of "refusals" and "no response" on days when taped and when not taped. During the 1971 population census of Morocco some of the work of 43 interviewers was taped.[5] The taped/not taped comparison was possible for only 30 interviewers, the work circumstances of the other 13 not being comparable. Of the 30 "comparables", 15 did less work on days when taped, 12 did more, and 3 did the same amount of work. It would appear that the taping had no effect on the length of the interview.

(MT 2) Tape recording improves the quality of the interview relatively to "true" facts.
In survey evaluation respondent's answers are sometimes compared with such records as family documents or employer's data. Provided the numbers of recorded and non-recorded interviews are large enough a macro comparison is then possible between the parameters obtained from the two (clearly, no micro or case-by-case comparison is possible). There is apparently little to choose between the validity of the recorded and not recorded interviews when compared with other validating data. Such distortions or losses as are found, appear to be related to social class: lower classes reply more accurately when taped, higher classes less accurately, at least in some British surveys (Belson, 1964; 1967).

(MT 3) Transcriptions, other forms of transformation from audial to optical record, and translations done twice by labour with same training disclose variable discrepancies without evidence of bias. High level labour is necessary to disclose biases.
Evidence available to date is uncertain and experiences are inconclusive, though the first part of the proposition is logically unassailable: there is no way to choose between two alternatives of equal quality.

The foregoing three hypotheses concern methodology of tape recording. The next nine concern the respondent-interviewer reactions that affect the methodology (M) of survey (S) taking. Some subtle hypotheses that would escape any attempt at quantifying from the tape are not tackled: e.g., the differences between the respondent's and interviewer's perceptions of the interest in and the nature of their interview.

(MS 1) The interviewer tends to conduct shorter and shorter interviews as the time of the day (particularly under trying climatic conditions) and the length of the interviewing campaign extends.
It is hypothesized that with time the interviewer:
becomes more impatient,
tends to ask leading questions,
interruptions not immediately relevant replies,
selects more understandable words,
adopts a more efficient mode of questioning.
Average daily production is a good indicator, but it cannot distinguish between the effects of boredom (the first two items) and the learning process (the last two items). The coding from tapes should differentiate (as it did not in Morocco) between features that indicate flagging "good behaviour" and features that suggest improvements in interviewer's competence. The interviewer can speed up the interview through the vertical method hypothesized in (MS 5). It appears that the vertical method arises as the day wears on, rather than with the dragging on of the campaign. Interviewers do start every morning with good intentions.

(MS 2) The type of the interviewer's interventions envisaged in hypothesis (MS 1) is more frequent with respondents that are illiterate, female, older or of lower socio-economic strata.[6]
There is little in the Moroccan data to affirm any such pervasive bias.

> "To be sure, one can hear in the tape recorded conversations impatient asides from a census-taker confronted with an illiterate woman who insists on a lower age for herself than the period of her declared residence in the city. A stammering old man's attempt to supply ages for his family are ignored by the (interviewer) and the child to whom he turns for answers. Some interviews include almost no questions concerning age, this information being calculated by the (interviewer) who reads the family civil paper. but such interviews are the exception, and evidence of pervasive cross-interview biases is weaker than expected" (Davis, 1973:47)

(MS 3) There is no difference in the mode of obtaining data and their reliability when answers are offered by one person only, by one person supported by documentary evidence, by several persons, particularly when each "self" replies for himself/herself.
Extreme instance can arise, though not on the Moroccan and Gambian tapes, when a whole village assists at the interview and objects loudly and chorally to the claim of an elderly female respondent to be still on the reproductive side of menopause. More generally a "self" takes more time to give information about himself than about others. In Morocco in 1971 questions concerning age (3.1 vs 2.5), birth place (5.5 vs 3.6), migration (4.8 vs 3.8), length of residence

(2.7 vs 2.1) provoked all more utterances in respect of the responding head of the household than for his spouse (Davis, 1973, tables B3 through B10). Whether this is due to the initial difficulty of understanding the question when posed for the first time or due to there being more information available to the respondent about himself (with consequent greater reliability of self-answered questions?) cannot yet be said.

(MS 4) The questions are not asked as formulated because of the tendency to slip, particularly in the case of answers, into generally prevailing and known definitions, no matter what the "official" survey definitions.
High marks can be given in this respect to taped interviews.[7] An anthropologist working on these data (Susan Davis) has several interesting examples in this regard.

(MS 5) To ask the same question for all members of the household (vertically) is faster than all the questions for each member separately (horizontally).
High marks to taped interviews. For the Morocan data the hypothesis has been confirmed for the age question and for the migration question (Davis, 1973). However, the faster, vertical method is liable to confusion of answers between members of the household. For a relevant finding see hypothesis (MS 1).

(MS 6) There is no objective way to differentiate between replies to questions that were not understood and where the answer was unknown to the respondent.
Subjective impressions can be formulated by a listener, but he longs to have an audio-visual taping to check out his impressions. No ideas on quantification, from transcriptions alone, have arisen.

(MS 7) The path from the initial response (eh?) through all the intermediaries to the finally recorded answer is a tortuous one and the number of utterances in between is inversely proportionate to characteristics listed in hypothesis (MS 2).
High marks can be given to taped interviews for their detective ability in this respect, but real benefits accrue only under special circumstances. If the distribution of female respondents is bi-modal -some shy in the presence of strangers, some more talkative than male respondents - the fact may be interesting, but cannot be easily used to improve the operations of the next survey. If the length in the interview varies with illiteracy of respondents who are evenly distributed over the country (Davis, 1973:35), nothing can be done about it. But if illiteracy is concentrated in certain strata, then it should enter significantly into the cost functions and influence sampling design. Strata by themselves were not a useful variable in Morocco (Davis, 1973: 49).

(MS 8) The use of a foreign language, literary rather than colloquial vernacular, and superior or educated accents increases the number of utterances during an interview.
High marks go to taped interviews in this respect provided transcriptions have been suitably coded. Whether the more subtle question of rapport and understanding between the two parties can be equally assessed is less certain.

(MS 9) Number of lines on transcription vary with socio-economic data and they also differ with the type of question.
Results of hypothesis (A1) and (MIG 2) indicate that interviewing in lower socio-economic strata (villages and shanty towns) is more expensive than in higher strata. However, strata are compounded with literacy. For survey design the finding on strata can be used, but not that on literacy. High marks go to taped interviews because of their ability to measure the "cost" of each question in terms of time or in terms of space on the transcription. The extraordinary similarity between the number of lines taken by the different subjects in table 1, except for household composition in hypothesis (C2), is remarkable for such different anthropological cultures as Gambia and Morocco.

To estimate completeness of coverage of a survey, taped interviews are a weak instrument, but at least two hypotheses can be suggested.

(C1) Under no circumstances is there ever an indication that whole dwellings are missed, but there is a slight possibility that other households within a given dwelling are being discussed in a given household in such a manner that indications of missed households are obtained. This is so especially, when definitional problems arise in multi-household dwellings and/or multi-family households.
There is no information that this possibility has been used.

(C2) Questioning about the whereabouts of age and sex groups particularly vulnerable to omissions is less intense (shorter in duration) than the importance of these groups (their predilection to be omitted from the age structure) would justify. Visitors are also enquired about less than the de facto and de jure problems would justify.
This possibility has not been tested. The question on household composition in Morocco, unusually expensive (57 lines in table 1) as compared with Gambia (17 lines), invites further exploration. The other relevant question, called in the table "family relations" called for 9 and 10 lines on the Moroccan and Gambian transcriptions respectively.

For the substantive discussion we selected three topics: age distribution, migration and occupation. Some results are available from Morocco and Gambia with regard to the first two. A thorough analysis would code utterances concerning age for use of documents (separating documents issued at the time of vital events

from documents completed retrospectively), the use of historical calendars, the family history method, the community comparative method, eye estimation, confident knowledge of respondents. There was no fertility in the Moroccan questionnaire and only few fertility questions in Gambia. Consequently, no experience in this all important field is available similar, e.g., to the embarassement reported on pregnancies, when still invisible, in Niger (Pool & Pool, 1971) and the consequent possibility of underreporting.

(A1) The age estimation provokes exchanges for heads of households, particularly with characteristics of hypothesis (MS 2).
High marks. In Morocco confirmation has been obtained for all the stipulations, except for the socio-economic strata. The contribution of strata is compounded with the other variables. However, this hypothesis is useful for sampling design, because one can "get at" these other variables through geographic strata.

(A2) The age estimation provokes more exchanges for households with documents and for members of lower birth order (younger children).
High marks. Confirmation has been obtained in Morocco for both stipulations. It would thus appear that consulting documents adds to the expense, but it also is likely to add to the accuracy, except when the document itself is doubtful (Krotki, 1973).

(A3) The likelihood that the interviewer uses leading questions when enquiring about age is greater when the head of household or speaker has characteristics of hyp. (MS 2).
High marks go to taped interviews for their ability to detect questions inviting agreement.

(A4) The likelihood of use of historical reference dates is greater when head of household or speaker has characteristics of hyp (MS2).
High marks. There is an inclination among respondents in Morocco to prop their memories with references to personal events rather than societal events. This inclination was shared by respondents in Niger (Pool & Pool, 1971). The ratio of personal to societal reminescences is higher the higher the proportions of the variables mentioned in the hypothesis (Davis, 1973: 40). For a slightly different experience in migration see hyp. (MIG 3). As suspected by the proponents of the historical calendar (Scott & Sabagh, 1970: 106) the link between the two series is weak.

(A5) The likelihood of broad and facile estimation of period of birth, of obvious guessing, of mistakes in mental arithmetics is greater when the head of household or speaker has characteristics of hypothesis (MS 2).
A hypothesis postulated "since Plato", but still of uncertain operationality.

(A6) Estimation of ages in decades or round numbers is more common for speakers with characteristics of hypothesis (MS 2).
One does not need taped interviews when the respondent is recorded on the questionnaire. When civil registration documents were originally issued on the basis of retrospective estimation the rounding disappears (deceptively). Other deceptiveness arises when interviewers calculate the age, repeat it aloud and implore the respondent to remember it from now on (Davis, 1973:31).

(MIG.1) Migration questions take up a high proportion of interview exchanges and time.
The migration questions appear to be expensive in both surveys on table 1, even in Gambia where they were limited to the place of birth and tribe.

(MIG.2) The number of exchanges per migration question(s) is greater for respondents with characteristics of hyp. (MS 2).
High marks. In Morocco the complexity of questions other than on migration, was felt more keenly in the shanty towns than in the rural douars[9] (Davis, 1973: 41). It cannot be said whether that would have been the experience with migration questions (if migration questions were asked from the rural questionnaire, which they were not) in view of the presumably simpler migrational history of the country people.

(MIG.3) The migration question "How long ago did you leave ..." is more easily answered if asked "How old (big) were you when ..."
High marks. The Moroccan tapes detect a tendency to switch to the unofficial form of the question (Davis, 1973: 47).

(MIG.4) The timing of migration questions caused more utterances than the geographic dimension of the moves.
Potentially high marks. Somewhat inconsistently with the outcome of hypothesis (MIG.2) and contrary to the expectation in hypothesis (A.3), there were more personal references than societal references on the timing of migration among the literate group than among the illiterates (Davis, 1973: 38).

(MIG.5) Controlling for variables in hyp. (MS.2), there are more leading questions for migration than for other subjects.
Potentially high marks. The hypotheses is based on the thought that the length and complexity of the migration questions in Morocco were an additional inducement to cut corners.

(MIG.6) Incomplete definitions concerning the urban-rural dichotomy and the "first city lived in" provoke much discussion.
Potentially high marks. It remains to be seen whether the general hypothesis (MIG.5) can be separated from the specific hypothesis (MIG.6). The operational consequences of the two hypotheses would be different.

(OCC.1) The number of exchanges between the interviewer and the respondent concerning occupations, is a function of whether or not the respondent is the subject.
Potentially high marks.

Table 1. - Lengths of exchanges between interviewer and respondent during the Gambian survey of 1972 and the Moroccan census of 1971

SUBJECT (1)	Question numbers G (2)	Question numbers M (3)	Average lines G (4)	Average lines M (5)	Average time - M min. (6)	Average time - M sec.
Household composition	a	U2	17	57	2'	53"
Family relations	b,c	U4-U8	10	9		28"
Age	d,e	U9-U10	23	21	1'	15"
Migration	f,g	U12-U16	21	48	2'	44"
Education	h,i	U22-U25	15	14		43"
Employment	l,m,n o,p	U28-U31	21	17	1'	25"
Irrelevant exchanges			13	13		45"
Non-comparable			34	117	10'	13"
TOTAL			154	296	16'	54"
Age as % of total			15%	7%		7%
Migration as % of total			14%	16%		16%

The table is based on 62 interviews from Gambia and 25 urban interviews out of 34 measured out of 800 taped interviews in Morocco; the "non-comparable" questions in the last line are mainly literacy, fertility and housing[8]

(OCC.2) Questions concerning unemployment produce less unemployment reported in the aggregate that the residual unemployment arising out of questions on employment.
Asking where and when the subject worked seems to draw away attention from unemployment. Labour force participation can, probably, not be studied without taped interviews.

(OCC.3) Serendipitous benefits of questions on occupations are lost in analysis.
Exchanges concerning occupations results in little occupational detail, but give considerable insight into the quality of life as perceived by the respondents, the impact of God Almighty, the method and chances of gathering a bountiful harvest, the fertility, location and distribution of the farmer's land. These details could become inputs into an aggregated subjective social indicator on quality of life rather than the construction of a meaningful occupational distribution. Taped interviews are a potentially valuable instrument in this regard.

(OCC.4) The richness of detail on occupations varies with the socio-economic level of respondents and subjects.
It is not clear a priori whether the variation is direct or inverse and whether it occurs with both the respondent and subject or only one.

(OCC.5) Whether in a developed society or underdeveloped, female respondents tend to overassess the occupation of their menfolk towards occupations with greater prestige.
Can be done only with some kind of follow-up and case-by-case check. Provided the respondent is recorded on the questionnaire there is no need for taped interviews in this respect.

It will be seen from the above summary of suggestions that the record of past achievements of tape recorded interviews is so far rather modest. Their further potentialities are mixed. Much of the disadvantage is not inherent in the method, but in the difficult circumstances in which it is being tried. As a minimum, however, the possibility remains that with a continuing development a new, even if only additional, instrument of objective evaluation will become available. The role in refining data collection procedures is more certain.

FOOTNOTES

[1] An attempt to short-circuit the transcriptions of the 1971 Moroccan interviews through a form with mere check listings, resulted in eight hours work for the migration and age questions alone on an average questionnaire, not very much less than the transcription of a whole questionnaire (Krótki, 1973, table 2). For more general introductions to the Moroccan data see Krótki, 1972; Krótki & Quandt, 1972; Quandt 1972.

[2] For example, the counting of transcription lines might give the same measure of "cost" as watch-measured timing (Krótki, 1973). For confirmation see table 1. The less expensive line counting requires some readiness to standardize transcriptions. Inter-language comparisons can be deceptive: French texts are typically one third longer than English, while texts in aglutinative languages can be still shorter.

[3] The thirty hypotheses listed in this paper rests on the work of researchers who in the early months of 1972 were engaged on the Moroccan data: Mohammed Abzahd, Mohammed Ayyad, Douglas Davis, Susan Davis, Karol P. Krótki III Dona MacLaren, Anna Quandt. For earlier ideas on audio-taping credit goes to John Blacker, Christopher Scott, William Seltzer.

[4] Matching in this case means aggregate matching of two groups (experimental and control) on as many characteristics as possible. It is not the matching required for record linkage and related procedures.

[5] For facilitating the tape recording in Morocco thanks are due to Tayeb Bencheikh, the then Director of Statistics, Abdelmalek Cherkaoui, his successor, Abdessattar Elamrani-Jamal, the Census Director, and Mohammed Rachidi, the Director of the Demographic Centre.

[6] For the Moroccan data socio-economic strata were approximated by geographic data. When we say "lower socio-economic strata" we mean simply "villages and shanty towns".

[7] If social surveyors use "official" definitions that are different from "native" ones, they break a fundamental rule of survey taking.

[8] David C. Roberts, the Government Statistician of Gambia made generously available to several researchers sets of transcriptions of interviews from the pilot census. Thanks are due to Don Peirce, a graduate student at the University of Alberta, for work on the materials available to the author.

[9] Douar is an Arabic word, used in Morocco to describe villages, or tribal areas, or more generally areas inhabited by country people.

REFERENCES

Belson, 1964 - William A., "Readership research in Britain". Business Review, 6 November.

Belson, 1967 - William A., "Tape recording: its effects on accuracy of response in survey interviews". Journal of Marketing Research, 4:253-260, August 1967.

Brody et al., 1951 - E.B., Richard Newman, and F.C. Redlich, "Sound recording and the problem of evidence in psychiatry". Science 113:379-80, April 1951.

Bucher et al., 1956a - Rue, C.E. Fritz, and E.L. Quarantelle, "Tape recorded interviews, some field and data processing problems." Public Opinion Quarterly 20:427-439, Summer Review.

Davis, 1973 - Douglas, Some effects on census interview content of interviewer and respondent characteristics. Haverford, Pa: Haverford College. Typescript.

Eitzen, 1952 - D.D., "Objective recording procedures in counselling and research". Marriage and Family Living 14:225-8, August.

Kantner & Zelnik, 1969 - John F. and Melvin, "United States: exploratory studies of Negro family formation - common conceptions about birth control". Studies in Family Planning 47:10-13, November 1969.

Krótki, 1972 - Karol J., "Programme d'exploitation de l'exercise EREB/CETI". Note d'Avis No. 21, 20 May 1972. Rabat, Maroc: Centre de recherche et d'etudes demographiques.

Krótki, 1973 - Karol J., Audiotaping of interviews for evaluation of social surveys. Annual meeting of the Statistical Sciences Association of Canada. Kingston, Ontario: Queen's University, June 1973. Reprinted as Discussion Paper No. ..., Population Research Laboratory, University of Alberta.

Krótki & Quandt, 1972 - Karol J. and Anna, The CETI/EREB exercise. Programme of analysis. Census evaluation by taped interviews. Rabat, Maroc: Centre de recherche et d'etudes demographiques.

Pool & Pool, 1971 - Janet E. and D.I., "The use of tape recorders to ascertain response errors in a KAP survey, Niger, West Africa." Annual meeting of the Population Association of America, Washington, D.C., April 1971.

Quandt, 1972 - Anna, The Collection of tape-recorded interviews during the 1971 census of Morocco. Document Technique du C.E.R.E.D., No. 4E. Rabat, Maroc: Centre de recherche et d'etudes demographiques, May 1972.

Scott & Sabagh, 1970 - Christopher and Georges, "The historical calendar as a method of estimating age: the experience of the Moroccan multi-purpose sample survey of 1961-63". Population Studies 24(1):93-109.

SMALL SAMPLE COMPARISONS OF CHI-SQUARE STATISTICS

Kinley Larntz, University of Minnesota

Several statistics are commonly used to judge the goodness-of-fit for counted data models. In this paper, two of these statistics will be compared with respect to their samll sample properties under the null hypothesis. The usual chi-square statistic (Pearson statistic) is defined by

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \qquad (1)$$

A suggested alternative statistic that has some asymptotically optimal properties is the likelihood-ratio statistic

$$G^2 = 2 \sum_{\text{all cells}} \text{Observed} \log_e\left(\frac{\text{Observed}}{\text{Expected}}\right) \qquad (2)$$

Many statisticians prefer the use of one or the other of these statistics, although among everyday users the Pearson statistic is far more popular. Also, some statisticians follow the practice of reporting both statistics (see for example, Goodman [1973]), but little guidance is available concerning the occurrence of large discrepancies between the two statistics.

THE MODEL

Comparisons between the statistics are made for a particular parametric model that arises naturally in a group helping situation. Individuals or groups are given the opportunity to help another individual in distress. The degree of help is graded I, II, or III: I for not helping, III for actively helping, and II for an intermediate action. Further details can be found in Fienberg and Larntz [1971] or Staub [1970]. Similar models are also used in component testing problems (see Easterling and Prairie [1971]).

Data were gathered for individuals and groups of size two. Let p_1, p_2, and p_3 be the probabilities of observing an individual with help graded I, II, and III, respectively. Then if the individuals in a group act independently and if only the higher grade of help is scored, p_1^2, $p_2^2 + 2p_1p_2$, and $p_3^2 + 2p_1p_3 + 2p_2p_3$ are the respective probabilities of observing I, II, and III for groups of size two.

Suppose N_1 individuals and N_2 groups are tested. Under the above assumptions, (n_{11},n_{21},n_{31}) follows a multinomial distribution with probability vector (p_1,p_2,p_3), and (n_{12},n_{22},n_{32}) follows a multinomial distribution with probability vector (g_1,g_2,g_3) where

$$g_1 = p_1^2$$

$$g_2 = p_2^2 + 2p_1p_2$$

$$g_3 = p_3^2 + 2p_1p_3 + 2p_2p_3 \quad . \qquad (3)$$

For this case the unique maximum likelihood estimates for (p_1,p_2,p_3) can be written down directly as

$$p_1 = (-n_{31} + \sqrt{n_{31}^2 + 4ac}\,)/2a$$

$$p_2 = r\hat{p}_1 \qquad (4)$$

$$p_3 = 1 - (1+r)\hat{p}_1$$

where

$$r = \frac{n_{21} - 2n_{11} - 4n_{12} + \sqrt{s}}{2(n_{11} + 2n_{12})}$$

$$s = (2n_{11} + 4n_{12} - n_{21})^2 + 8(n_{21} + n_{22})(n_{11} + 2n_{12}) \qquad (5)$$

$$a = (1 + r)[(n_{11} + 2n_{12})(1 + r) + (n_{31} + 2n_{32}) + 2n_{22}(1 + r)/(2 + r)] \quad , \qquad (6)$$

and

$$c = n_{11} + 2n_{12} + 2n_{22}/(2 + r) \quad .$$

The selection of this model for making comparisons between the likelihood-ratio and Pearson chi-squares provides several advantages:

(a) The model depends on two parameters, p_1 and p_2, and thus the goodness-of-fit test for the null hypothesis involves the estimation of these parameters. Comparisons are therefore made for a composite null hypothesis.

(b) Since the maximum likelihood estimates can be written down in closed form, iteration is not necessary for finding the estimates. This is important when considering the feasibility of doing large amounts of computation.

(c) Examining (3), note that the probability of Help Grade I for groups is p_1^2. When p_1 is small, p_1^2 is quite small. Thus the selection of this model allows for comparisons of very skew multinomials, which means comparisons can be made for small as well as moderate minimum cell expectations. Previous studies (Cochran [1952], Yarnold [1970]) have indicated that, for small expected values, the Pearson statistic does not follow the chi-square distribution well, while some suggestion has been indicated (cf. Bliss [1967]) that the likelihood-ratio statistic would be

better in such situations.

SMALL SAMPLE PROPOERTIES UNDER THE NULL HYPOTHESIS

Under the null hypothesis the goodness-of-fit statistics, X^2 and G^2, have asymptotic chi-square distributions with 2 degrees of freedom. However, for small samples the chi-square approximation in many cases does not agree well with the actual distribution. Several studies (Cochran [195], Fisher [1958], Roscoe and Byars [1971], Yarnold [1970]) have given conflicting points of view as to at what point the approximation is "reasonable" for the Pearson chi-square statistic. Standard rules specify that the minimum cell expectation should be 5, with possibly a few smaller. The emphasis here will not be on finding such a rule, but in comparing the likelihood-ratio and Pearson statistics with regard to the approximation. In other words, we ask for small samples, which of the two statistics is better approximated by the asymptotic chi-square distribution?

The initial task in this study of the small sample properties is to determine the distribution of the statistics G^2 and X^2 when the null hypothesis holds. Several methods are available to handle such a problem. The principal method used here was that of enumeration. The number of possible outcomes of two trinomials with sample sizes N_I and N_G is given by

$$\text{Outcomes} = \binom{N_I + 2}{2} \binom{N_G + 2}{2} \qquad (8)$$

For $N_I = N_G = 8$, the number of possible outcomes is 2025. Thus, for a given value of (p_1,p_2,p_3), N_I, and N_G, the distribution of G^2 and X^2 were determined by computer.

One question that arises in the use of this method is how to deal with zero cell counts and zero expected values. The maximum likelihood estimates were extended by continuity to provided well-defined procedures. In the same manner, when a cell had zero expected value, it contributed zero to the chi-square statistic.

Figure A gives a contour plot of the mean of G^2 for $N_I = N_G = 8$. Barycentric coordinates were chosen to represent the 3 probabilities. Each corner of the triangle represents one of the probability vectors (1, 0, 0), (0, 1, 0), and (0, 0, 1), while a general point in the triangle corresponds to the probability vector (p_1,p_2,p_3). Figure B gives a similar plot for X^2. The asymptotic mean for both statistics is, of course, 2.0. The mean of G^2 overshoots that value for a large set of (p_1,p_2,p_3). The peak value is approximately 2.51. In viewing Figure B, it can be seen that the mean of the Pearson statistic is a smoother function of (p_1,p_2,p_3) than the mean of G^2. For a large set of (p_1,p_2,p_3), the mean of X^2 is close to 2.0. The peak value is approximately 2.12. Thus, considering the mean only, the Pearson statistic appears better.

Another method of comparison is to check the agreement of the actual small sample percentage points with the corresponding asymptotic values. Results analogous to the case of the mean hold here. Namely, the likelihood-ratio tends to overshoot the corresponding large sample value while the Pearson statistic tends to be closer to the asymptotic value for a large range of (p_1,p_2,p_3).

Several questions concerning the likelihood-ratio statistic arise from these results. First, is it still possible that the optimality properties of the likelihood-ratio statistic carry over in spite of the poor characteristics of its null distribution? This will be the subject of another paper comparing the powers of the statistics. Second, can the statistics be easily adjusted to remove some of its poor behavior? And third, exactly how does the likelihood-ratio behave as the "small" sample size increases? An attempt at answering the last question will be given below.

The question of adjusting the likelihood-ratio statistic poses large difficulties. A simple-minded correction for the mean yielded mixed results, partly due to a problem of overcorrection with respect to size. Other corrections involving more moments or quantiles may be possible, but practical use would require a simple multiplicative or additive correction, such as those given in Bartlett [1947] and Box [1949].

PROPERTIES OF THE LIKELIHOOD_RATIO CHI-SQUARE STATISTIC

The asymptotic distribution of G^2 for the model considered here is that of a chi-square variate with 2 degrees of freedom. Figure C gives a graph of the mean values of G^2 for (.6, .2, .2). Figure C is indicative of what happens to the mean as the sample size changes. It begins below its asymptotic value, rises to a peak, and descends to the correct value. The true sizes follow a similar pattern. Because of the discreteness of the distribution, the rise and descent may be slightly rocky, but the general pattern remains the same.

The sample size at which the peak is reached varies considerably depending on the probability vector (p_1, p_2, p_3). Some evidence has been given that the minimum cell expectation governs the closeness of the small sample distribution to asymptotic theory for several chi-square problems (see for example, Cochran [1952], Cramer [1946], Odoroff [1970], Yarnold [1970]). In the problem at hand, small expected values are found for small values of p_1 (since the first cell for pairs has probability p_1^2) and for very small values of p_2 and p_3. Evidence from this study indicates that the larger minimum cell expectation cases are closer to the behavior predicted by the asymptotic theory.

POWER CHARACTERISTICS

In order to compare the power functions of the test statistics, it was necessary to adjust for the level of significance differences between X^2 and G^2. Let the adjusted level be defined as

$$\text{Adj. level } (z) = \sup_{(p_1,p_2,p_3)} P(\text{statistic} \geq z) \quad (9)$$

Thus for a given alternative $(p_1, p_2, p_3; g_1, g_2, g_3)$, the power of X^2 or G^2 can be computed as a function of the adjusted level.

Many methods can be used to compare the power functions of the statistics. One interesting comparison can be made by means of the median significance level (Joiner [1969]. For a particular alternative, let

$$\text{M.S.L.} = \text{Adjusted Level } (z_M) \quad (10)$$

where z_M is the median of the statistic under the alternative distribution. In comparing two statistics, the one with the lower median significance level would be considered better. For this and several other methods of comparison, it was found that the Pearson statistic was more powerful than the likelihood-ratio for most alternatives.

The stochastic limit ratio (defined below) gives a method of determining the alternatives where the likelihood-ratio was more powerful than the Pearson. For an alternative $p_a = (p_1, p_2, p_3; g_1, g_2, g_3)$, let $G^2(p_a)$ be the value of the likelihood-ratio chi-square calculated using $n_{11} = p_1$, $n_{12} = p_2$, $n_{13} = p_3$, $n_{21} = g_1$, $n_{22} = g_2$, and $n_{23} = g_3$. Similarly, let $X^2(p_a)$ be defined. Then for an alternative define the stochastic limit ratio as

$$\text{S.L.R.}(p_a) = \frac{G^2(p_a)}{X^2(p_a)} \quad (11)$$

When S.L.R. is large (71.05), the likelihood-ratio statistic appears more powerful based on small samples; whereas, with S.L.R. > 1, the Pearson statistic is better. The differentiation in the middle range (1 - 1.05) is not clear with some cases going to likelihhod-ratio and some to Pearson. However, for large areas of the alternative parameter space, S.L.R. < 1.

CONCLUSIONS

For one special model with a composite null hypothesis, the samll sample distributions of two chi-square statistics were examined. Using as criterion the closeness of small sample dsitribution to the asymptotic chi-square approximation, the Pearson chi-square statistic is by far the more desirable. The likelihood-ratio statistic has an expected value in excess of the nominal and yields far too many rejections under the null distribution.

It was also noted that the expected value and level of significance for the likelihood-ratio statistic displayed a consistent regularity in which the mean and level rose to a peak and then declined toward the asymptotic value as the sample size increased.

Power comparisons also indicated the desirability of using the Pearson statistic over the likelihood-ratio -- at least when proper adjustments are made for the differing levels of significance.

REFERENCES

Bartlett, M. S., "Multivariate Analysis," Journal of the Royal Statistical Society, Series B, 9 (1947), 176-197.

Bliss, C. I., Statistics in Biology, Volume I, New York: McGraw-Hill Book Co., 1967.

Box, G. E. P., "A General Distribution Theory for a Class of Likelihood Criteria," Biometrika, 36 (1949), 317-346.

Cochran, W. G., "The χ^2 Test of Goodness of Fit," Annals of Mathematical Statistics, 23 (1952), 315-346.

Cramer, H., Mathematical Methods of Statistics, Princeton, N. J.: Princeton University Press, 1945.

Easterling, R. G. and Prairie R. R., "Combining Component and System Information," Technometrics, 13 (1971), 271-280.

Fienberg, S. E. and Larntz, K., "Some Models for Individual-Group Comparisons and Group Behavior," Psychometrika, 36 (1971), 349-367.

Fisher, R. A., Statistical Methods for Research Workers, 13th ed., New York: Hafner Publishing Co., 1958.

Goodman, L. A., "Guided and Unguided Methods for Selecting Models for a Set of T Multidimensional Contingency Tables," Journal of the American Statistical Association, 68 (1973), 165-175.

Joiner, B. L., "The Median Significance Level and Other Small Sample Measures of Test Efficacy," Journal of the American Statistical Association, 64 (1969), 971-985.

Odoroff, C. L., "A Comparison of Minimum Logit Chi-Square Estimation and Maximum Likelihood Estimation in 2 x 2 x 2 and 3 x 2 x 2 Contingency Tables: Tests for Interactions," Journal of the American Statistical Association, 65 (1970) 1617-1631.

Roscoe, J. T. and Byars, J. A., "Sample Size Restraints Commonly Imposed on the Use of the Chi-Square Statistic," Journal of the American Statistical Association, 66 (1971), 755-759.

Staub, E., "A Child in Distress: The Influence of Age and Number of Witnesses on Children's Attempts to Help," Journal of Personality and Social Psychology, 14 (1970), 130-140.

Yarnold, J. K., "The Minimum Expectation in X^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution," Journal of the American Statistical Association, 65 (1970), 864-886.

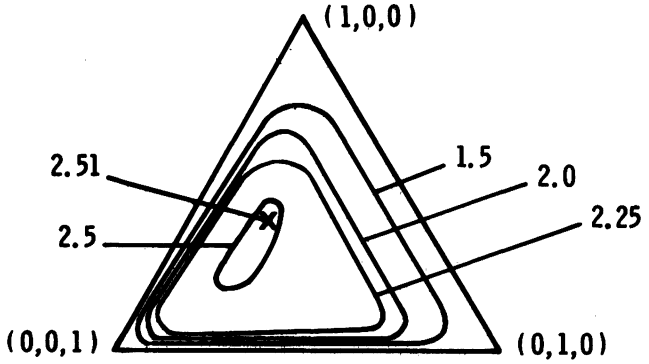


Fig. A

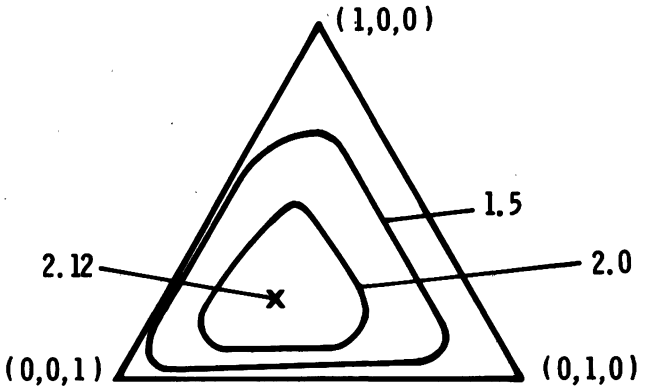


Fig. B

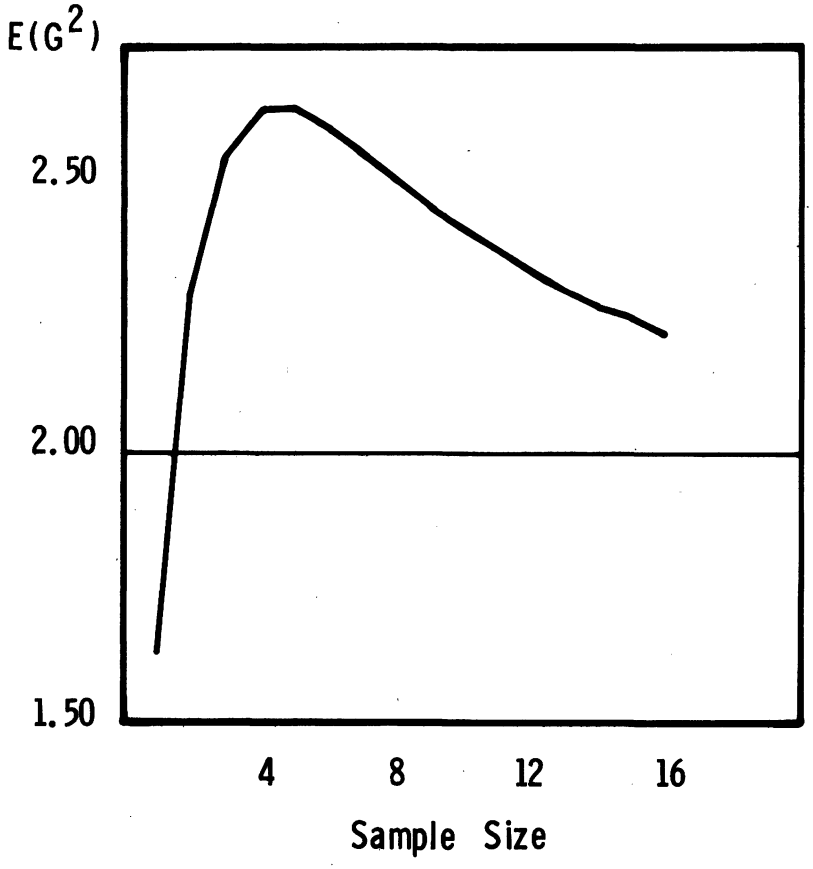


Fig. C

THE ECONOMICS OF HOSPITAL UTILIZATION UNDER INSURANCE
Some Preliminary Findings

Irving Leveson and Regina Reibstein, New York City Department of Health

For the last fifteen years the analysis of medical markets has been dominated by "Roemer's Law," the proposition that the supply of hospital beds determines the demand for inpatient care. There is a great deal of evidence that the number of patient days of care is roughly proportional to the number of short-term general hospital beds.

A number of hypotheses as to the nature of the relationship are considered and their theoretical and empirical bases are explored. The material presented here represents an abbreviated portion of the new empirical evidence which is developed. A complete copy of the paper is available upon request.

Most of the available evidence relates to market-wide changes. We were concerned that this might mask important developments taking place at the individual hospital level. The approach chosen was to examine hospitals undergoing expansion in order to determine why some hospitals are able to fill their beds and others are not. The sample consists of all cases of hospitals with an initial size of 400 beds or more which increased in size by 50 beds or more during the 1960's. The findings presented apply to 39 hospitals in Standard Metropolitan Statistical Areas operated by nonprofit corporations. This category excludes hospitals controlled by religious organizations. Six observations were excluded because of missing data. The mean size of hospitals was 593 beds and the mean increase 125 beds.

The Model and the Data

The dependent variable is the elasticity of census with respect to beds. The percentage change in census per percentage change in beds is generally given by $\eta=\frac{\Delta C}{\Delta B}/\frac{C}{B}$, so that the elasticity is equal to the ratio of the marginal occupancy rate to the average occupancy rate. A value of 1 indicates that the occupancy rate of the added beds is equal to the occupancy rate prior to expansion. In order to avoid distortions due to unequal expansions, the arc is actually used.

The arc elasticity of census with respect to beds is calculated from the period two years prior to expansion until two years after construction of the new beds is completed. The possibility of going until the third year was rejected because the mean change in census from the second to the third year was only one percent for all hospitals, there was more of an opportunity for extraneous factors to have an influence and the sample size would have been reduced. The mean elasticity for 39 nonprofit corporation hospitals was .95 and the standard deviation was .44. The variables and hypotheses follow. Except as noted, the area specific variables apply to SMSA's.

Price

The design of this study is especially suited to determining whether price is important at all. The extensiveness of insurance coverage, absence of consumer information and role of technical factors in the medical care decisions make it necessary to first raise the question in this form. Hospitals undergoing expansion are generally a close substitute for other hospitals in their area so that the effect of price on demand for the individual hospital can be expected to be far greater than the effect of price on the quantity of medical care consumed.

The measure used is the difference between the average annual percentage change in the hospital's expense per patient day from two years before construction until two years after and the average change for similar hospitals in the community. The geographic unit is the self-designated city reported in the American Hospital Association's "Guide Issues." Price changes during construction are heavily influenced by the cost of paying off the mortgage and therefore reflect the lavishness of the project, construction costs, interest rates, and the extent of philanthropy and thus can be considered exogenous to bed utilization. A greater increase in costs is expected to result in a lower elasticity of census with respect to beds.

An alternative formulation treats expense per patient day of the study hospital and of comparable hospitals in the area as separate variables. The term $b(X_1-X_2)$ is replaced by $b_1X_1-b_2X_2$ in order to determine if $b_1>b_2$. One hypothesis is that consumers considering care at the study hospital will have only very incomplete information as to prices at other hospitals and as a result will respond less sensitively to changes in prices at other hospitals. Another possibility is that new facilities bear a disproportionate share of the adjustment to proportional cost increases in the system as a whole. This would appear as a greater negative response to a hospital's own price increase than the positive response to the increase in the average price at other hospitals. Such behavior could arise because of the effects of habit formation on the ability of a new hospital to attract either patients or staff.

Insurance

The higher the level of hospital insurance coverage in the area, the more readily beds are expected to be filled since the average out-of-pocket cost for the patient will tend to be lower. Surgical insurance reduces the cost of treatment and is expected to be associated with greater utilization as well. The percentage of the population covered by hospital and surgical insurance were each constructed by a)estimating

enrollment in private insurance in the SMSA for the population under age 65, adjusting the state average for the average difference between large SMSA's and other areas derived from a regression of city size on insurance coverage across states, b) adding in average coverage for the aged, and c) adding in coverage for the indigent by applying the SMSA's proportion of the state's poor to state Medicaid enrollment. Aggregates of states were used where SMSA's were located in more than one state.

The correlation between hospital and surgical enrollment rates in the nonprofit corporation samples was .84. Results are shown using hospital coverage alone. In an alternative specification, surgical coverage is used as a measure of the percentage of the population with both hospital and surgical coverage and a separate variable for the excess of hospital coverage over surgical coverage is used as a measure of the proportion of the population with hospital coverage only.

These variables test the actual effect of insurance coverage on utilization. A test was also made of the effects of expectations on the part of the hospitals as to the impact of insurance. If hospitals expected Medicare and Medicaid to make it easier to fill beds but in fact beds were not filled then we would expect to find lower elasticities among hospitals for which the decision to construct was made after these programs were anticipated. Hospitals completed in 1969 or later are assumed to have been committed in 1965 or later and are coded one in a dummy variable to measure the effect of expectations.

Hypotheses Relating to Beds and Population

Level and Change in Hospital Size. Hospitals would be more likely to reduce occupancy if increasing size were associated with greater specialization and less likely if increasing size presented a greater opportunity to average risks. The greater the expansion in beds, the lower the elasticity if there are long- or short-run difficulties in absorption of additional beds. The absolute size and change in beds are used.

Number and Growth in Beds Per Thousand Population. Beds per capita in the SMSA provide a direct test of whether there is a tendency for marginal occupancy rates to be lower where bed supply is greatest. The Community's (AHA) rate of growth of beds per capita during the period two years before construction until two years after was expected to be negatively associated with elasticity if there is a tendency toward saturation.

Population and Population Density. A larger population, given density and beds per capita, is expected to result in greater occupancy because of the larger potential market. A denser population for a given population size is expected to be associated with higher occupancy because of the proximity to markets. Data are in millions and thousands per square mile respectively.

Population Growth. The larger the percentage change in population from 1960 to 1970, the greater the demand for hospital care. High rates of growth may be anticipated, however. Past growth may be a measure of expected future growth. Hospitals may be more likely to build beds which will not be filled immediately in areas with rapid expected growth in future demand, resulting in lower elasticities.

Physician Availability

The larger the number of physicians in the area per hospital bed (1970), the greater the ease of staffing. In an alternative formulation, the impact of specialists per bed and general practitioners per bed is examined separately in view of recent evidence that specialist supply tends to be associated with greater hospitalization and general practitioner supply with less. Physician availability is also measured by the percentage growth in physicians per bed in the state between 1963 and 1970. State data are used deliberately in order to avoid an identification problem stemming from the effect of the hospital expansion itself on the demand for physicians in the SMSA.

Teaching Status

Dummy variables were created for three teaching categories: medical school affiliation, limited graduate training and non-teaching. The choice of which variable to omit was made by entering the dummy variables in a stepwise regression in order to select that combination of dummy variable categories which maximized the contribution to R^2.

Findings

The results of the linear ordinary least squares regression across 39 nonprofit corporation hospitals are shown in Table 1. The coefficient of determination is .35. The only variable which is significant at a high level is price. The simple R^2 between price and the elasticity is .09.

A 1 percent increase in the average annual rate of price change from two years before construction until two years after is associated with a .052 decrease in the elasticity of census with respect to beds. The standard deviation of price is 4.1 so that a one standard deviation variation in price is associated with a .21 point difference in elasticity or approximately one half a standard deviation. The coefficient of the hospital insurance variable is sizable and in the expected direction but not statistically significant. The coefficient of beds per capita is positive contrary to expectations, but is not significant while bed growth is far from significant.

In the second equation, the price components are examined separately. The hospital's own price change is highly significant with a coefficient of .06 and a standard error of .023. However, the price change of other hospitals, while positive, has a coefficient of only .03 and a

t ratio less than one. While there is no significant difference between the absolute values of the coefficients of the price variables, the size of the difference is substantial.

The third equation examines the alternative formulation of insurance coverage. The coefficient and significance levels for the combination of hospital and surgical coverage are similar to the values for the hospital coverage variable. The variable for additional hospital coverage over surgical has no effect.

Finally, the effects of general practitioners per bed and specialists per bed are examined separately. There is still no significant effect of physician availability in the area on the change in occupancy of hospitals which expand.

A test was made to determine whether there is any tendency for hospitals with a low elasticity to have smaller price increases in the future relative to other hospitals as a means of raising occupancy. In this analysis the elasticity is the independent variable and price the dependent variable. The average and change in price relative to other hospitals is taken from the year two years after hospital expansion, the final date for the elasticity, until the fourth year after expansion. Capacity changes took place early enough for the follow-up period data to be available in 32 of the hospitals. The relationship was estimated as

$$P=-5.40+4.15\eta \qquad R^2=.10.$$
$$(2.27)$$

The hypothesis is accepted at the .05 level on a one sided test. The coefficient is substantial indicating an 8 percent price difference after two years for a one percent difference in elasticity. The earlier estimates indicate that a 4 percent per year difference in price could be expected to have a large effect on utilization. We have only looked at price differences which are realized, of course, and an infrequent use of price adjustment could arise from a tendency for other hospitals to follow price changes, together with low response of market demand to price.

Discussion

The evidence of the importance of price as a modifier of the response of census to beds is consistent with the findings of studies which depend heavily on insurance as a price determinant. The finding that a one standard deviation change in price is associated with a half standard deviation change in the utilization elasticity invites comparison with Martin Feldstein's evidence that half of the response of census to beds is a pure availability effect and half a price effect.[1] The interpretation is very different, however.

Less than one-tenth of the variation in the response of census to beds among hospitals is associated with price. The mean price change is -.2, virtually zero. The mean elasticity of census with respect to beds, the ratio of the marginal to the average occupancy rate is .95, essentially indistinguishable from 1. While we find evidence that price is an important variable even when represented by hospital expenses per patient day, the tendency is for the average response of census to beds to hover around one even after price effects are removed.

Price may also play some role in adjustments to excess capacity, as the post construction comparisons suggest. But it must be noted that when a 500 bed hospital with 90 percent occupancy expands to 600 beds, the difference between an elasticity of .5 and an elasticity of 1 is one of whether the hospital will end up with an occupancy rate of 82.5 percent or of 90 percent, and differences of that magnitude are able to persist. We have found no evidence to cast doubt on the proposition that when hospital beds expand the marginal occupancy rate is about equal to the average.

An expanding hospital may pull some patients from each of several others, and we have made no tests of the effects of hospital capacity changes in the occupancy rate in the entire market. To some extent this was tested in reverse by looking for effects of changes in capacity of other hospitals on use of the study hospitals. The lack of significant effect is consistent with Sander Kellman's finding that capacity changes have only a small effect on occupancy.[2] The lack of a negative impact of beds per capita also is indicative of little or no tendency for the marginal occupancy rate to be declining in response to capacity changes.

The findings appear to provide further support for the notion of a systematic relationship of census to beds. Proper interpretation requires consideration of explanations beyond the scope of this part of the analysis.

FOOTNOTES

1. Martin Feldstein, "Hospital Cost Inflation, A Study of Nonprofit Price Dynamics," American Economic Review, 61, No. 5 (December 1971), pp. 853-872.

2. Sander Kellman, "Utilization of and Investment in U.S. Short-Term Hospitals," unpublished Ph.D. dissertation, University of Michigan, 1970.

TABLE 1

REGRESSION ANALYSIS OF ARC ELASTICITY OF CENSUS WITH RESPECT TO BEDS

	Basic Equation	Price Components	Insurance Components	Physician Components
Intercept	-1.610	-1.286	-1.371	-1.612
Price Difference	- .053		- .052	- .053
	(.021)		(.020)	(.023)
Hospital Insurance	.022	.022		.022
	(.014)	(.014)		(.015)
Expected Insurance	.234	.261	.186	.235
	(.200)	(.179)	(.178)	(.207)
Size of Hospital	.001	.001	.001	.001
	(.001)	(.001)	(.001)	(.001)
Change in Hospital Size	- .001	**	- .001	- .001
	(.001)	**	(.001)	(.001)
Beds per Capita	.136	.124	.145	.137
	(.084)	(.080)	(.082)	(.087)
Bed Growth	- .020	- .022	- .025	- .020
	(.037)	(.023)	(.034)	(.038)
Population	.010	*	*	.010
	(.065)			(.068)
Density	.032	.028	.040	.032
	(.057)	(.055)	(.056)	(.062)
Population Growth	.009	.008	.011	.009
	(.009)	(.009)	(.009)	(.009)
Physicians per Bed	.878	- .671	- .971	
	(.863)	(.858)	(.857)	
Physician per Bed Growth	8.320	8.586	7.588	8.343
	(6.874)	(6.198)	(6.145)	(7.078)
Medical School Affiliation	- .200	- .195	- .462	- .201
	(.216)	(.212)	(.275)	(.224)
Graduate Affiliation			- .258	.314
			(.258)	(.281)
Non-teaching Hospital	.314	.188		
	(.274)	(.294)		
Hospital Price		- .060		
		(.023)		
Area Price		.030		
		(.037)		
Hospital and Surgical Insurance			.023	
			(.014)	
Hospital Insurance Only			.013	
			(.022)	
General Practitioners per Bed				- .979
				(4.011)
Specialists per Bed				- .865
				(1.003)
R^2	.350	.362	.358	.350

NOTE: Standard errors are in parentheses.

* Did not enter under F = .001.

** Less than .0005

OUT-OF-POCKET EXPENSE FOR HEALTH CARE BY LOW INCOME FAMILIES

Regina Loewenstein, Columbia University

Introduction

From November 1968 through June 1969 a nationwide interview survey was conducted with a sample of low income families in low income areas to study the effects of Title XIX, Grants to States for Medical Assistance Programs, called Medicaid, that was passed in 1965. (1) This talk will discuss the health care experiences of persons in the survey who lived in the 39 states that had Medicaid programs on January 1, 1969. (2)

The interviews included questions about utilization of health services, out-of-pocket expense for these services, needs for health care, knowledge of Medicaid, and practical and attitudinal barriers to receipt of health care. Comparisons and interrelationships of these indices among three groups of low income persons in states with Medicaid programs will be used as indicators of the effects of this program. The three groups, defined by Medicaid eligibility laws in each state at the time of the study, were public assistance recipients, medically needy persons and Medicaid ineligibles.

At the time of the interviews there were 39 states with Medicaid programs. All public assistance recipients in these 39 states were eligible for Medicaid. In 16 of these 39 states the Medicaid law also covered some persons not receiving and/or not eligible for public assistance but satisfying criteria for medical assistance with regard to age, income, family size, family structure and assets. These persons were called medically needy. The low income persons in the 39 states who were not eligible for public assistance or Medicaid were called Medicaid-ineligibles.

In 1969, the federal Medicaid law required that all states cover inpatient care and medical care in institutions, clinics and from private doctors. The Medicaid program in many states also paid for dental care, prescribed drugs, and other health services.

Methodology

The definition of low income families to be interviewed was designed to maximize the number of families eligible for Medicaid and to include families with incomes slightly above the criteria for Medicaid in each state. (3) Since the limit of federal assistance in Medicaid programs was related to payments for Aid to Families with Dependent Children (called AFDC), the income criteria for inclusion in this study was less than 150 percent of AFDC payments in each state for the specific family size.

In order to increase the likelihood of finding low income families and thus reduce the costs of screening to locate families eligible for interviews, the first stage of sampling was a random selection of 400 enumeration districts from those with the highest proportions of low income families in strata defined by state and size of community. Within each of these 400 enumeration districts, households were randomly selected for screening questionnaires which asked about family size, public assistance and amount of income in four weeks. Families receiving public assistance or with incomes below 150 percent of AFDC payment for their size in their state were called survey eligible and were interviewed. In summary, the sample for the entire study was a multi-stage stratified cluster sample of low income families in low income areas in conterminous United States. (4)

About 15,000 screenings were completed, of which about one-third were eligible for interviews. Interviews were completed with 5,382 families. Eighty percent (or 4,277) of these families lived in the 39 states with Medicaid programs. Since each person in each low income family was covered in the interviews, data about the 12,309 low income persons in these families will be discussed.

Considering the Medicaid laws of each state with regard to public assistance, income, age, size of family, family structure and assets, each of the 4,277 families and thus each of 12,309 persons was classified according to Medicaid eligibility. About 39 percent of the low income persons covered in the interviews were public assistance recipients, 22 percent were medically needy persons and 39 percent were not eligible for Medicaid coverage in their state.

Because age and sex distributions of the three eligibility groups were not the same, consideration was given to using age-sex adjusted indices instead of unweighted indices about all persons. For each of ten of the most important dependent variables, patterns of differences among the three Medicaid eligibility groups were the same for unweighted indices as for age-sex adjusted indices.

Data about five age-sex groups in each Medicaid eligibility group are in the detailed report. (5) For almost every variable, the patterns of differences of the three Medicaid eligibility groups were the same for all ages and for each age-sex group. Therefore, unweighted indices of persons of all ages will be used in this talk to compare the three Medicaid eligibility groups.

Demographic Characteristics

The differences in demographic characteristics of the three Medicaid eligibility groups reflect the regulations for federally aided programs for public assistance and for medical assistance. The majority of families receiving public assistance consists of a mother and one or more children under 18. In this study, the

public assistance recipients had the significantly highest proportion of persons under 18 and the largest percentage of families consisting of a female head with children. (6)

Since most Medicaid programs covered needy persons 65 and over, many of whom were not receiving public assistance, almost one-third of the medically needy persons were 65 and over in contrast with 9-12 percent of the two other groups. The Medicaid ineligibles had the most intact families and the highest percentage with employed heads of families.

Also, the public assistance recipients had twice as many Black persons as the two other groups, more persons from large cities and almost no one with health insurance. The medically needy group had the most small families and the most families with annual income below $2,000.(7) The Medicaid ineligibles had the highest percentage of heads of families who were high school graduates.

Utilization of Hospitals

Public assistance recipients had higher rates of hospital stays than the other two groups, even though the medically needy persons had a much higher proportion of persons 65 and older. The higher hospital rate for public assistance recipients was not due to stays for pregnancy. Among women 18-44, public assistance recipients had a considerably higher rate of stays without pregnancy than the other two groups and had about the same rate of stays with pregnancy as Medicaid ineligibles.

Almost three-fourths of the stays of public assistance recipients were without charges to the person, family, Medicare and/or insurance, in contrast with 20-25 percent for the other two groups. (8) About 60-70 percent of the stays of medically needy persons and Medicaid ineligibles were paid for by Medicare and/or insurance.

The National Center for Health Statistics (NCHS) presented data from their 1968 survey about persons in families with income under $5,000 who did and did not receive public assistance. (9) In both surveys the proportions of persons with one or more short-stay hospital episodes were higher for persons receiving public assistance than low income persons not receiving aid. The values of these percentages for persons within each age group who did and did not receive aid were almost the same in the two studies.

Among persons 45 and older the proportions of persons with stays were higher for medically needy persons than for Medicaid ineligibles, and the proportions of the Medicaid ineligibles were the same as of persons without public assistance in the NCHS study.

Ambulatory Medical Care

Among the three groups, public assistance recipients had the highest rates of total ambulatory medical care and of ambulatory medical care without charges and the lowest rate of visits with charges. Medically needy persons and Medicaid ineligibles had similar rates of total visits and of visits with charges, but the medically needy persons had slightly more visits without charges than Medicaid ineligibles. More than 80 percent of the visits for public assistance recipients were without charges, in contrast with one-third for medically needy persons and one-fourth for Medicaid ineligibles.

Mean total visits and mean visits to clinics per year were highest for public assistance recipients, but mean visits to private doctors were about the same for the three groups. The proportions of visits without charges were higher for clinic visits than doctor visits in each group. Only one-fourth of doctor visits by medically needy persons were without charges, even though all medical visits by them were legally covered by Medicaid.

Another publication based on the National Health Interview Survey showed that the use of physicians in 1969 was about the same for persons with incomes under $3,000 and $3,000 to $3,999; these two income groups correspond to the income of almost all of the families in the Columbia study. (10) Compared with this nationwide sample of persons from families with incomes under $4,000, the public assistance recipients 18 and over had more ambulatory medical care, the medically needy persons and Medicaid ineligibles 18 and older had less care, and persons under 18 in each Medicaid eligibility group had less ambulatory medical care.

In both the Columbia and NCHS studies, the mean physician visits per year for each age group were higher for public assistance recipients than for other low income persons. (9)

The relationships of utilization of health services to the need for health care, barriers, and knowledge of Medicaid eligibility will now be discussed.

Needs for Health Care

For each age-sex group, public assistance recipients had the highest proportions with poor health, with health poorer than persons of the same age, with some disability days, and with long-term health problems. Within each eligibility group, persons with the poorest health, as indicated in these various ways, had the most medical care. But within each category describing health status, public assistance recipients had the most care, and the other two groups had about the same amount of care.

Barriers

Twenty questions were asked about financial, practical and attitudinal barriers to health care. The public assistance recipients had the highest proportions of families with nine of these twenty barriers, the medically needy persons were highest on two barriers, and the Medicaid ineligibles were highest with regard to two barriers.

The barriers expressed most often by the respondents for the families of public assistance recipients were: prefer to use different facilities, have doubts about doctors, get out of bed too soon when ill, know of trouble in getting health care because of discrimination, persons on welfare get poorer care, have no telephone, traveled one hour or more and/or waited one hour or more for a visit for medical care, and were bothered by waiting in a clinic or office. The barriers to health care expressed most frequently by medically needy persons were: understanding one's health better than doctors and expectation of illness when old. Medicaid ineligibles most frequently said that they wait to see a doctor when ill and that the head of family loses pay if he takes a day off for health reasons.

Relationships to utilization of three of these barriers were analyzed. These questions were about: not calling a doctor right away when ill, doubts about doctors, and understanding one's health better than doctors. Respondents who expressed each of these barriers had lower rates of ambulatory medical care than persons who gave the opposite response, but within each of the two categories of response to each of the three questions, public assistance recipients had higher utilization rates than the other two groups of low income persons.

Thus, although public assistance recipients had the highest proportions of families with almost half of the barriers, they had higher utilization rates than medically needy persons and Medicaid ineligibles among both those who did and did not express these barriers to health care.

Within groups of persons in the same categories about need or about barriers, public assistance recipients had more health care than Medicaid ineligibles for whom free care was rarely available, and more than medically needy persons who less frequently knew about their Medicaid coverage.

Knowledge about Medicaid

More than one-half of the public assistance recipients and only one-fourth of the medically needy families knew that their state had a Medicaid program and that they were eligible for it. Also, public assistance recipients more often correctly knew what services were covered by the Medicaid program in their state.

For both public assistance recipients and medically needy persons, utilization rates of covered services without charges were compared among four groups of persons defined by combinations of correct knowledge of the person's eligibility and correct knowledge of coverage of the specific service by the Medicaid program. Within each eligibility group, utilization of services without charges was highest for persons presumed eligible and with correct knowledge of coverage of that service. Within each of the four knowledge groups, utilization rates were higher for public assistance recipients than for medically needy persons.

Higher utilization rates by persons who knew they were eligible indicate that utilization and knowledge were related. But it is not possible to determine from these data what proportion of persons had knowledge because of their needs and what proportion sought care to be paid by Medicaid because they knew about the Medicaid program and their eligibility for it.

The welfare worker was the major source of information about the Medicaid program for public assistance recipients who have frequent contacts with their social workers. Public media and other individuals were the major sources of knowledge about Medicaid for the medically needy persons.

Out-of-Pocket Expense

Out-of-pocket expense was defined as the amount of money paid by the person or family for health services after payments were made by Medicaid, Medicare and insurance companies. Persons eligible for Medicaid paid for a service covered by Medicaid if they were not aware of their eligibility for Medicaid and what it covered, if they neglected to or were reluctant to apply for Medicaid coverage, if papers about eligibility were being processed, or if the provider of care did not participate in the Medicaid program. Participation indicated that the provider accepted the fee scale, was willing to wait for payment, and was willing to submit required reports.

Although all Medicaid programs covered medical services, about 8 percent of public assistance recipients and almost half of the medically needy persons who had ambulatory medical care in four weeks had some out-of-pocket expense for it. Almost 40 percent of Medicaid ineligibles who had ambulatory care had no out-of-pocket expense for it because of coverage by Medicare and/or insurance, free care in clinics or by doctors, and unpaid bills.

Hospital bills were also covered by Medicaid in all states. Of those with hospital stays in one year about one-tenth of public assistance recipients, more than one-third of medically needy persons, and almost half of Medicaid ineligibles had some out-of-pocket expense for hospital bills. The mean out-of-pocket expense per person with expense was $100 - $200 per year for hospital bills and $100 - $150 for in-hospital medical care.

About 10 percent of public assistance recipients, 25 percent of medically needy persons and 30 percent of Medicaid ineligibles had some out-of-pocket expense for one or more health purposes in four weeks. The means of total out-of-pocket expense per person in four weeks were $1, $5 and $8, respectively, for the three groups. The total mean expense and the means for each of five categories of expense were significantly higher for Medicaid ineligibles than for medically needy persons, and also higher for medically needy persons than for public assistance recipients.

Almost half of the out-of-pocket expense for

public assistance recipients was for prescribed medicines. For medically needy persons, about one-third of the out-of-pocket expense was for medical care and another one-third was for prescribed medicines. For Medicaid ineligibles almost half of the out-of-pocket expense was for medical care. Hospital bills were almost one-tenth of the out-of-pocket expense for public assistance recipients and about one-sixth for the other two groups.

The financial hardships of some low income persons were reflected in the range of out-of-pocket expense for all health care. Out-of-pocket expense for all health care was $30 or more in four weeks for one percent of public assistance recipients, three percent of medically needy persons and almost five percent of Medicaid ineligibles.

The means of out-of-pocket expense for all health care in four weeks per person with some expense were $13, $18 and $26, respectively, for the three groups. When persons had expenses for prescribed medicines the means for the four-week period were about $10 in each group, and for persons with expense for dental care the mean per person was about $25 in each group.

Summary

Three groups of low income families in low income areas in Medicaid states have been compared with regard to utilization of, out-of-pocket expense for, need for and barriers to health care plus knowledge of Medicaid coverage.

Public assistance recipients eligible for Medicaid had the highest needs for health care, and in spite of the high frequency of many barriers to care, they had the highest utilization rates. Because of their eligibility for and knowledge of Medicaid coverage they had the lowest out-of-pocket expenses.

In each age-sex group, medically needy persons expressed less needs for health care than public assistance recipients. Among persons with similar needs, medically needy persons had less health care than public assistance recipients although they expressed fewer barriers. Lower utilization by them may reflect not only less knowledge about the Medicaid program but also more pride and reluctance to apply for assistance for which they were eligible. Some Medicaid eligible persons may have felt that the difficulties to learn how to apply for medical assistance did not warrant the financial help that would be received. In addition, some medically needy persons may have had difficulties in finding private doctors who were willing to take Medicaid patients.

Medicaid ineligibles - that is, poor persons with incomes slightly above the criteria for Medicaid eligibility in their states - had lower needs for health care, and utilization rates similar to those of medically needy persons. Their out-of-pocket expense was higher than that of persons eligible for Medicaid because of the lack of coverage by Medicaid, inability to get free care, and no or inadequate health insurance coverage.

Implications

Two of the program implications from this study were as follows. First, the demands for services to be paid by Medicaid increased as more public assistance recipients and especially medically needy persons learned about the program, their eligibility for it, and the services in some states.

Second, persons with incomes slightly above the criteria for Medicaid eligibility have less care than Medicaid eligibles with similar needs for health care because of financial hardships. Programs are needed to help these low income families receive health care.

Plans for future programs and legislation would be facilitated by further studies of health care of low income families. Examples of questions to be studied are: quality of care given by providers; satisfaction of consumers; factors that contribute to over and under-utilization of health care; and the effects of changes in Medicaid programs in specific states with regard to eligibility requirements, covered services, efforts to inform eligible persons about Medicaid, and control of quality of care.

FOOTNOTES

(1) The research on which this talk is based was done in 1967-1972 by Columbia University School of Public Health under contract from Social and Rehabilitation Service, Department of Health, Education and Welfare. The contract numbers were WA-406, SRS-ORDT-68-01 and SRS-69-50. Drs. Otto Reid and Oliver C. Moles were the members of the SRS staff who worked with the Columbia staff. Dr. David Wallace of Columbia worked on the study in the planning stages. Interviews were conducted by Audits and Surveys, Inc. Victor Soland was responsible for data-processing operations, and Dr. Andre Varmá was statistical consultant.

(2) The ten states in conterminous United States that did not have Medicaid programs on January 1, 1969 were: Alabama, Arizona, Arkansas, Florida, Indiana, Mississippi, New Jersey, North Carolina, Tennessee and Virginia. The data in this report are based on interviews done in other states in conterminous United States and in the District of Columbia.

(3) To be consistent with the Medicaid law, a family was defined as a married couple or one adult with or without children under 21, or an unrelated individual 21 or over. For example, if a household had parents with two children under 21 and one child over 21, the first family would be the couple and the two youngest children and the second family would be the oldest child alone.

(4) The sample design was developed by Lester

Frankel, Audits and Surveys, Inc. with members of the staffs of Social and Rehabilitation Service and of Columbia University School of Public Health.

(5) Loewenstein, Regina. Effect of Medicaid on Health Care of Low Income Families. Two volumes. Columbia University School of Public Health, 1971.

(6) All differences between groups discussed in the text were tested with consideration of the multi-stage sample design, and were significant on the five percent level.

(7) Income includes money from earnings, pensions, public assistance, and other sources.

(8) Services were classified as "with charges" if some charges had been incurred by the person, family, Medicare and/or insurance. When providers of services were paid by Medicaid or welfare agencies, the care was classified as "without charges." That is, the concept of charges was with regard to the consumer and not to society as a whole. Prepaid services, services covered by Medicare plus Medicaid, and services completely covered by insurance were classified as with charges but with no out-of-pocket expense to the person or family.

(9) Health Characteristics of Low Income Persons. DHEW Publication No. (HSM) 73-1500. Series 10, Number 74. July 1972.

(10) Age Patterns in Medical Care, Illness, and Disability, United States, 1968-1969. DHEW Publication No. (HSM) 72-1026. Series 10, Number 70. April 1972.

MODEL BUILDING - RACE AND IQ
A MODEL-BUILDING ANALYSIS OF THE JENSEN HYPOTHESIS

James M. Lucas - E. I. du Pont de Nemours & Company

In his article "How much can we boost IQ and scholastic achievements?" (1) Arthur Jensen implies that the observed difference in IQ test scores between the black population and the white population is due primarily to genetic effects. Discussants of this paper have stated that determining the genetic and environmental contributions to this difference is a difficult problem (2). With the presently available techniques and the available data, this not a difficult problem - it is an impossible problem.

IQ studies show that the black population's IQ test scores are about 15 points lower than the test scores of the white population. Even when the populations are matched for environmental variables such as socioeconomic status, the black population scores lower, though the differences are only about half as great (3). There is no doubt that differences in IQ test scores exist; the only question is to their cause.

We will follow a regression model-building approach to study this question. We will show what happens when we try to estimate genetic effects and certain environmental effects. Let us first examine the problem of estimating the effects of genetic differences between a black population and a white population. The data consist of IQ measurements and other measurements such as socioeconomic indices for the individuals being studied. The effects of genetic differences between the two populations can be modeled by an indicator variable which is assigned one of two values - for example, zero for a member of the white population and one for a member of the black population. A regression model can be fitted to the data containing this indicator variable by the method of least squares. The regression coefficient calculated for the indicator variable can be considered to be an estimate of the effect of genetic differences between the two populations. The regression coefficient for the indicator variable will be approximately -15 if it is the only variable in the model; it will be about half as large if environmental variables such as socioeconomic status are included. The above statements summarize the studies that have been run to this time (3).

Now consider the problem of estimating IQ difference caused by certain environmental differences between the black population and the white population. We will consider modeling the total effect of such environmental variables as caste, prejudice, and differential expectations. The difference in IQ test scores between the black population and the white population caused by these environmental variables can be modeled by an indicator variable which has two values - for example, zero for a member of the white population and one for a member of the black population. The regression coefficient calculated for the indicator variable, in this case, can be considered to be an estimate of the sum of the effects of such environmental variables as caste, prejudice, and differential expectations.

For both the genetic model and the environmental model the indicator variable will take on exactly the same values, and the calculated regression coefficient will be exactly the same. That is, if the indicator variable, which in the environmental model accounts for the sum of the effects of certain environmental variables, is the only variable in the model, its regression coefficient will be about -15. If other environmental variables are included, the regression coefficient for the indicator environmental variable will be about half as large.

The effects of the environmental variables accounted for by the indicator variable and the effects of genetic differences accounted for by the indicator variable are exactly "confounded". There is no way of disentangling the relative contributions of genetic variables and the effect of certain environmental variables on observed differences in IQ test scores between the black population and the white population. We can measure the sum of the contributions of the genetic and environmental variables, but we cannot measure either the genetic effects or the environmental effects separately.

Because of the confounding of genetic effects and certain environmental effects, the following three mutually contradictory hypotheses (among others) are consistent with the observed data:

Hypothesis A:

The genetic difference between the black population and the white population causes about a 15-point difference in IQ test scores. The effect of environmental variables on IQ test scores is negligible.

Hypothesis B:

Environmental variables tend to depress

the black population's IQ test scores by about 15 points. The genetic effect is negligible.

Hypothesis C:

The environmental variables tend to depress the black population's IQ test scores by about 30 points. The genetic variable favors the black population by about 15 points.

Generally, some combination of Hypotheses A and B is believed; however, because of the confounding, Hypothesis C is not rejected by the data.

The Problem of Confounded Data

The above problem is caused purely by the kind of data that can be obtained. There is no statistical technique that can resolve the problem without different kinds of data. Standard regression "corrections" for environmental variables such as socioeconomic status give a better analysis than an analysis that makes no environmental correction; however, such an analysis cannot correct for environmental effects due to variables such as caste, prejudice, and differential expectations.

We can design a "thought" experiment that would enable us to measure unambiguously the contribution of environmental and genetic differences to the observed difference in IQ test scores between the black population and the white population. The experiment will take 2N families; N black families and N white families. The families should have at least two children. Pair a black family and a white family whose children are born on the same day and switch children at birth. The parents should not be told of the switch. The white children should receive injections to make them appear black, and the black children treatments to make them appear white. With such an experiment the genetic contribution and the environmental contribution to observed IQ test differences could be estimated unambiguously. Such an experiment is clearly impossible in our society.

Indirect Approaches

Since it is impossible to find the reason for differences in IQ test scores between the black population and the white population directly, indirect approaches have been tried. One approach is to attempt to discredit studies that show environmental effects due to caste, prejudice, or differential expectations. Discrediting these studies would indicate that these environmental effects cause small, if any, changes in IQ test scores. Jensen (1) used this approach in his article when he attacked the study by Rosenthal and Jacobson, "Pygmalion in the Classroom" (4).

Finding areas where the environmental effects are small is a second indirect approach. Kleinberg (5) cites a study by Clark (6), of Los Angeles, on Negroes in 1923. In five elementary schools the average black IQ test score was 104.7, while for all elementary pupils the average IQ test score was 106. Clark indicated that he felt the IQ scores were about 5% too high but that there were no significant differences in IQ test scores between the black population and the white population. Clark's study indicates that in parts of southern California in the 1920s there were only small IQ and environmental differences between the black population and the white population. In the same article Kleinberg, who lived in Canada for 25 years, expressed doubts that Tanser (7), in his Canadian study, had really found an area where environmental effects were small.

Tanser found IQ differences of around 15 points between the black population and the white population in an area where there was little overt prejudice. Table XXIII from Tanser is reproduced as Table I.

TABLE I

Median IQs and Percentiles of White and Negro Pupils on the National Intelligence Test According to Rural or Urban Environment

		Whites	Negroes
No. Pupils	Urban	339	43
	Rural	47	60
Median IQ	Urban	104.68	89.08
	Rural	96.29	90.06
Median Percentile	Urban	57.16	27.0
	Rural	40.75	27.83

In the Urban population the blacks were of a lower socioeconomic class than the whites. The rural populations consisted mainly of farmers who owned their own farms, so the blacks and whites were much more closely matched in background. Shuey (3) excused the poor performance of rural whites. She explained that most of the rural whites had migrated to Canada within the last two generations, so their children were not as familiar with standard English as were urban whites. Recent linguistic work (8) has shown that blacks learn an English that is different from standard English. This tends to lower their IQ test scores, since most IQ tests are based on standard English. Thus, in Table XXIII the IQ scores of all blacks, as well as rural whites, are probably depressed from their true

potential scores. Tanser's table does show that in matched populations, such as the rural black-white population, IQ test scores differences do tend to be significantly smaller than 15 points.

Since the difference in IQ test scores between the black population and the white population may contain both environmental effects and genetic effects, a single large study, such as Clark's, which found less than a two-point spread in IQ test scores, can be used to place an upper limit on the possible genetic effect (9), while studies such as Tanser's, which find the usual difference in IQ test scores, indicate that the usual combination of environmental and genetic variables is having its effect.

Estimating Heritability

Estimating the genetic and environmental components of the observed difference in IQ test scores between the black population and the white population is impossible. Estimating the genetic components and the environmental components of the variability in IQ test scores in a white population (the only ones on which heritability studies have been made) (10) is difficult because of high correlations between environmental variables and genetic variables. With the available data, slight changes in the estimation technique can lead to large changes in the estimates. For example, using the data from Jensen, Light (11) obtained an estimate for heritability of .63 which is appreciably lower than Jensen's estimate of heritability of .75 which was obtained from the same data.

Light used a more complete breakdown of the environmental variability. He considered as a separate environmental component a covariance term which measured such things as a tendency of adoption agencies to match children to foster parents. Including this effect as part of the environment component reduced the estimate of the genetic component.

Heritability estimates apply only to the population being studied. The study by Burt and Howard (12), which estimated environmental and genetic components for a homogeneous white population (London school children), makes the disclaimer "we should like to insist on the very limited nature of the problems we have tried to solve. Neither here nor elsewhere have we attempted to reach any overall statement about the relative contributions of heredity and environment to mental efficiency or 'intelligence' as manifested in ordinary everyday life or among all classes and conditions of men." A bias in the Jensen article is the fact that he carefully makes a similar statement (p. 47), while on a different page he uses heritability estimates obtained from a small population for a larger population which would be expected to have more environmental variation (p. 36).

Since heritability is the variability that can be accounted for by genetic components divided by the total variability, large populations such as the total population of the US have a lower heritability than the heritability estimates obtained from the small homogeneous populations on which heritability studies have been made (13).

Combining the data from the subpopulation studies, which have been made in an attempt to obtain an overall heritability estimate for the population in the United States, involves many methodological difficulties. Jencks (14) discusses these in detail and obtains an estimate of heritability as .45 for the US population. The remaining variability is broken down into environmental effects (35%) and genetic-environmental interactions (20%). This is significantly lower than the .80 heritability usually claimed by Jensen. Jensen's estimate is an average of subpopulation estimates, so it cannot contain all the environmental variability.

Jensen's heritability discussion makes few explicit statements. The implication is that if heritability is very high, environmental effects are small; so observed differences in IQ test scores between the black population and the white population must be due to genetic effects. Small changes in heritability estimates can cause large changes in the inferences that are made. The difference in heritability between the two models -

a) Genetic effects cause all the difference in IQ test scores between the black population and the white,

b) Genetic effects cause none of the difference in IQ test scores between the black population and the white,

- is only 10%! This is shown in (15); it is discussed further in the next paragraph.

Since the proportion of blacks in the US population is small, the difference in the mean between the black population and the white population accounts for a small percentage of the total variance (which is about 225). Jensen [(2) page 81] states, "In terms of proportions of variance, if the number of Negroes and whites were equal, the differences between racial groups would account for 23% of the total variance, but - an important

point - the differences within groups would account for 77% of the total variance." Jensen's statement is misleading; it overestimates the variance caused by black-white differences. Since the black population is only 11.2% of the total population, the difference between racial group means accounts for only 10% of the total variance; thus, the differences within the black and white populations account for 90% of the total variance (15).

Related Results

We should note that a large heritability for IQ test scores does not imply that there will not be large differences in IQ test scores when environmental differences are large. Deutsch (16) examined the intra-pair differences for IQ test scores of identical twins reared apart. In the studies he cited the maximum within-pair difference ranges from 14 to 30 points. The IQ differences were correlated with differences in environment. Where large IQ differences were found, there tended to be large environmental differences, and large differences were rare when environmental differences were small. Reducing environmental deprivation can have good effects regardless of the heritability (16).

Wheeler (18) studied 3,000 Tennessee mountain children between the ages of 6 and 16 whose average IQ was 82 in 1930. After improvements in environment, the average IQ had increased to 92 by 1940. If one is to claim that most of the observed difference in IQ test scores between the black population and the white population is caused by genetic effects, one must argue that the isolation effects suffered by the Tennessee mountain children are much greater than the environmental effects that adversely affect the black population's IQ test scores.

Measuring Environmental Effects

The direct measurements of environmental effects discussed by Jensen (pp. 52-54) will tend to underestimate the environmental effects. The method used is to obtain environmental indices, then to perform a multiple regression using these indices as the independent variable with IQ measurements as the dependent variable (19). The variability accounted for by these indices is taken to be the environmental effect. Few researchers would claim that their indices could account for all the sources of environmental variability; thus, these studies tend to underestimate the variability in IQ test scores accounted for by the environment.

Conclusion

We have shown that the complete confounding of genetic effects and certain environmental effects makes it impossible to determine their relative contribution to observed differences in black-white IQ test scores. The type of data that can be obtained from large scale studies of IQ test scores cannot resolve the question of the cause of IQ test score differences between the black population and the white population. If such studies are to be made, they must be justified on other grounds.

Since black-white IQ test score differences account for only ten percent of the variability in IQ test scores, and since the heritability for IQ in the US population must be less than the .80 estimate Jensen obtains from subpopulation studies (Jencks .45 heritability estimate is more consistent with all the data), an environmental model for observed differences in black-white IQ test scores is very feasible. Policy decisions should not be governed by the assumption of a genetic cause for IQ differences between blacks and whites.

References

(1) Jensen, A.R. How much can we boost IQ and scholastic achievement? Harvard Educational Review, 1969, 39, 1-123

(2) Discussion: How much can we boost IQ and scholastic achievement? Harvard Educational Review, 1969, 39, 273-356

(3) Shuey, A.M. The Testing of Negro Intelligence, Social Science Press, New York 1966. This book summarizes most of the black-white IQ comparisons that have been made. Most of the studies summarized make no correction for environmental variables. Studies which correct for more than one environmental variable are rare.

(4) Rosenthal, R., & Jacobson, L. Pygmalion in the Classroom, New York: Holt, Rinehart, & Winston, 1968

(5) Kleinburg, O. Negro-White Differences in Intelligence Test Performance: A New Look at An Old Problem. Amer. Psychol., 1963, 18, 198-203

(6) Clark, W.W. Education Status of Los Angeles Negro Children, Dept. of Psychology and Educational Research, Los Angeles City Schools, 1923

(7) Tanser, H.A. The Settlement of Negroes in Kent County, Ontario.

Reprint 1970 Negro University Press, Westport, Conn.

(8) Dillard, J.W. Black English. Random House, New York, 1972

(9) Unless (a) Environmental effects favored Los Angeles blacks over Los Angeles whites, (b) there was selective migration of high IQ blacks to the Los Angeles area, or (c) Clark's study was invalid, a genetic difference in IQ test scores between the black population and white population of less than 5 points is indicated. For Clark's data the Z value for a difference in IQ test scores as large as 5 points is $Z=5.3=(5.0-1.3)/(15\sqrt{1/510+1/4326})$ when $\sigma=15$ is assumed for both populations. This gives a probability of about 10^{-7} (one in ten million) that the black-white difference in IQ test scores is greater than 5 points.

(10) Jensen states (p. 64) "... all the major heritability studies reported in the literature are based on samples of white European and North American populations, and our knowledge of the heritability of intelligence in different racial and cultural groups within these populations is nil".

(11) Light, R.J. Biometric issues in measuring the genetic component of human intelligence. The New York Statistician, 1971, 22(5), 3-8

(12) Burt, C., and Howard, M. The multifactorial theory of inheritance and its application to intelligence, Brit. J. of Stat. Psy., 1956, 8, 95-131

(13) Burt and Howard comment on this with respect to the less controversial (than IQ test scores) variable, stature of students. This variable has an extremely high heritability. On a population of University students and their relatives, Burt and Howard calculated the genetic component of stature as 97% leaving only 3% to be accounted for by nongenetic factors. They say: "These figures, of course, hold good only for the particular population studied, namely University students and their relatives. Data obtained from London school children indicate that in boroughs where (at the time of our earlier surveys) poverty and malnutrition were rife, the nongenetic variance might amount to nearly 20%." Heritability estimates must be reduced when generalizing from populations with small environmental effects to those with larger environmental effects.

(14) Jencks, Christopher. Inequality, Basic Books Inc., New York, 1972

(15) The proportion of the total variance accounted for by black-white IQ differences is $B W\Delta^2/\sigma^2$ where Δ is the difference in IQ test scores, σ is the standard deviation of IQ test scores, B is the proportion of blacks and $W=(1-B)$. When $\Delta=\sigma=15$ (the most commonly used value for Δ and σ), the above formula gives 25% for B=.50. This is slightly greater than Jensen's 23%. When the 1970 Census B=.112 is used, the above formula gives slightly less than 10% as the proportion of the total variance accounted for by black-white IQ test score differences!

(16) Deutsch, M. Happenings on the way back to the forum: Social Sciences IQ and race differences revisited. Harvard Educational Review, 1969, 39, 523-557

(17) Unless heritability is 1.0 so that the environment has no effect. Also see (13)

(18) Wheeler, L.R. A comparative study of the intelligence of East Tennessee mountain children. J. Educ. Psychol., 1942, 33, 321-334

(19) Usually only a linear term is entered in the model for each independent variable. Interaction and higher order terms do not enter the model. Thus, only the "additive" effects of environment are estimated. This is a second reason that the environmental effects are underestimated.

EMOTION, REASON, AND RISK*

Clifford J. Maloney, Bethesda, Maryland

What textbook of logic does not point out in its preface that formal logic deals, not with how men actually think, but with how they would think if their instrument of thought were reason? How men actually think is left to the science of psychology or to unfathomable intuition in the case of women. To my knowledge, no previous thinker has gone so far as to enunciate the fundamental principle: "Decision-making is an emotional act".

The twenty-five years of my professional life have been dominated by this phenomenon. This and the fact that I seem to see the same effect in many if not most other human endeavors makes it seem worthwhile to call the phenomenon to the attention of statisticians generally. Instances in which the outcome is determined by the dominance of the emotional setting in which decision-making occurs fall into four general types of consequences. First, incidences distinguishable only by their emotional content lead to differing results, whereas had the decision been a rational one the outcomes would have been similar. Differences in sentences for the same crime but by a different accused may be the most common example. Sensitivity of a decision to a change in circumstances is a second consequence of emotional decision-making. Over the same twenty-five year period the automobile and road construction were in the driver's seat, so to speak, until some time in the 1960's, when silent spring found a powerful voice; now seeming to be somewhat muted by the necessity to obtain new sources of energy and for fuller exploitation of existing supplies.

A third major characteristic of emotional decision-making is its extreme polarity. What is heinous for you to do is unavoidable, if regrettable, when I do it. It has often been remarked that there are no wars of aggression. All initiations of hostilities are but reactions to incidents, threats, or dangerous preparations of the attacked.

It is a fourth area of emotional decision-making that is my chief concern. Certain topics, actions, or devices per se are inordinately highly charged emotionally. Biological warfare is such a subject. Its stigma arises from two sources. It is perverted medicine and medicine is expected to be nice. It was placed under the Army Chemical Corps and Americans think that chemical warfare was a German innovation. No such stigma attaches to the tank, even the flame thrower, for the tank was invented by the English. I cannot claim to have examined all of the discussion, but I have not seen any that explains why it is better to be killed by a bullet or a bomb than by gas or a disease. That both gas and disease may be incapacitating and not permanently injurious is of course ignored in such discussions.

So emotional is medicine generally that no activity is permitted which could conceivably have contributed to the death or injury of a person whether intentionally, negligently, or otherwise. Only in medicine and health is perfect performance demanded. Military commanders may be censored for reckless exposure of their men but they are never expected to win battles without losses. Physicians, too, are not expected to practice medicine without occasionally losing a patient. But the loss is to be due to the disease or disability. Every act of the physician should universally have the consequence of reducing or minimizing the consequences; never, of itself, to contribute to the injury of the patient.

Vaccine administration does not meet this test. The first large scale attempt to protect a threatened population from an epidemic by immunization--actually inoculation--occurred in Boston in 1721. The medium was not a carefully attenuated strain but fully virulent smallpox. This first extensive application of the procedure was attended by a highly emotional dispute as to its legitimacy and its effectiveness which has continued to the present in connection with the later vaccination against smallpox. But in the process the standard of performance has grown inordinately, and the nature of risk has been dramatically altered. In 1721 few if any people could escape exposure to smallpox. Today few are likely to be exposed. The relevance of this example to the topic of this paper arises on the one hand (1) from this minuscule risk, (2) from the even smaller risk of a catastrophic outbreak, and (3) from the small but still appreciable risk of untoward results from vaccination itself. On the other hand we have a government which condones a known killer--tobacco. Both questions are discussed in largely emotional contexts. And the adopted actions in the two cases differ widely while it would appear that they should be similar.

The history of vaccine introduction, like the history of food and drug regulation, and indeed every form of endeavor affecting health and safety where explicit overt action is involved has continued to be highly emotional. The highly emotional reaction to death or injury traceable to explicit overt action in contrast to the tolerance for far greater consequences of neglect or indifference has often been noted by astute critics; without in general arousing the public. For example most modernizations of applicable law have followed some dramatic occurrence, often however far less costly in the aggregate than other less dramatic, more constant, threats to life and safety. This point was expressed at the first Academy Forum of the National Academy of Science on the Design of Policy on Drug and Food Additives held in Washington, D. C. on May 15, 1973 by Peter Hutt, Assistant General Counsel for Food and Drugs, DHEW, in the words:

> "In short public policy design and execution with respect to the safety of food and drugs is highly and perhaps irretrievably, controversial. It raises up a welter of subjective and emotional views that often obstruct rational analysis and that severely hinder regulation by scientific decision-making.

. . . . . .

* Scheduled but not delivered.

"One does not need a degree in science to hold and express deeply-felt beliefs on the degree of risk or uncertainty society should accept from food and drugs. Nor, indeed, does a scientific background equip one with any greater insight into the intricacies of this type of policy issue or any more impressive credentials or greater authority to act as an arbiter in resolving these matters. As long as we remain a free society, these basic philosophical principles will, and properly should, remain the subject of intense public scrutiny and debate.

. . . . . .

"As a lawyer, I am not only accustomed to the adversary process but also a strong advocate of it. Nevertheless we must be careful to prevent trial by combat from replacing reasoned decision-making on important safety issues".

The words "statistics" or "statistical decision-making" do not occur in Mr. Hutt's address. Either he is unaware that a discipline of statistical decision theory exists or he rejects its applicability in safety regulation of food and drugs. Moreover the Interdisciplinary Communications Program of the Smithsonian Institution has held or is planning some half dozen conferences on the Philosophy and Technology of Drug Assessment beginning in May of 1970. The roster of attendees for the first two conferences does not contain the name of even one professional statistician though I am informed that this omission was rectified by the third conference.

The status of decision-making in activities affecting life and safety then is: (1) little use is made of formal statistical techniques including statistical decision-making; (2) discussion in these areas is highly emotional and inconsistent. One thesis of this paper is that these two characteristics are related.

The best encapsulation of the impossibility of rational decision-making when emotion is dominant and why that I have seen is continued in a letter to the editor of Medical World News for November 26, 1965. The correspondent, Dr. John T. Flynn, referring to the opposition of "regular" physicians to practitioners of psychoanalysis, wrote: "The tragic fact is that our disrespect of psychiatry and psychological theories must remain a futile posture until a solid theory of human psychological functions can be established. Unhappily, attempts to refute psychoanalytic theory run aground upon the rigidity of human faith. When a solid basis of scientific fact does not underlie an understanding of some area of nature then pure faith and belief in a system of some kind seems a human necessity. It does little good to chip away at such an unsubstantiated system by means of appeals to logic, furious attacks, or sarcasm. The only way one can displace inadequate or fraudulent theory is by offering a superior substitute based upon clear scientific understanding. The age of reason still remains an age of faith".

As averred by the Assistant General Counsel of FDA, decision-making in the regulation of food and drugs has since the beginning and is daily an emotion-stirring process. Why is that? Dr. Flynn implies that we lack the appropriate scientific basis for a rational solution. But why is that? Mr. Hutt appears to assert that the situation will inevitably remain emotional; and in doing so handicaps rational decision-making. Dr. Flynn in an analogous discipline asserts that emotion can be eliminated so soon as an effective rational explication becomes available. It is his view that I subscribe to: That is: emotional discussion is a symptom of a lack of an adequate rational base. Its disruption of rational decision-making while real and destructive is secondary.

But Mr. Hutt has highlighted the essential lack in the safety field when he writes: ". . . there appears to be no public or scientific consensus today on the risk or uncertainty acceptable to justify the marketing of any substance as a food or drug". But, of course, it is agreed that if any risk whatever is acceptable it is a very low one, of the order of one in a 100,000 or even less.

I will assume here that this is in fact the one ingredient missing to permit removing most of the emotional content of decision-making in food and drug regulation and indeed in all activities involving human life and safety. Hence the problem arises because of the very low risks which are tolerable. The currently dominant Baysean approach to probability assessment arose for the most part in the context of the so-called unique incident probability estimation, the Amchitka underground test, where nothing comparable has ever occurred in the past.

In this latter context, but applicable to both, the author has supplied an approach to the numerical estimation of very low probabilities in a previous paper entitled "A Probability Approach to Catastrophic Threat" available from the National Technical Information Service. The effect of this approach is to replace human estimates of absolute probabilities by relative estimates.

NOTE: The thoughts and opinions expressed are exclusively those of the author.

OPTIMUM VALUES OF LIFE TABLE FUNCTIONS

S. Mitra, Emory University

1. The Problem

The object of this investigation is to isolate three important characteristics of a life table, namely;

1) the age at which the derivative of l_x, i.e., of the proportion surviving from age zero to age x (which is uniformly negative) assumes its maximum or negatively minimum value;

2) the age at which μ_x or the force of mortality - $(dl_x/dx)/l_x$ is minimum, and;

3) the age at which the expectation of life $\overset{0}{e}_x$ is maximum.

It is well known that because of relatively large value of infant mortality compared with those of early childhood ages, the derivative of l_x is large and negative at age zero. Thereafter, the derivative continues to increase till it attains a maximum at some age (usually in the age interval 10-14), and declines thereafter till l_x becomes zero. Correspondingly, the force of mortality assumes a relatively large value at age zero, begins to decrease thereafter until a minimum is reached (usually around age 12) and continues to increase and becomes quite large at the end of the age span. The pattern of variation of $\overset{0}{e}_x$ is not that apparent, and as has been shown earlier (Mitra, 1971), the maximum value of $\overset{0}{e}_x$ is reached after age zero but usually before age 5, the proof of which will be repeated in this paper for the sake of continuity.

2. Maximum value of $\overset{0}{e}_x$

By definition, $\overset{0}{e}_x = T_x/l_x$ where $T_x = \int_0^{\alpha-x} l_{x+t}\,dt$, α being the upper age limit, so that $l_\alpha = 0$. Thus,

$$\frac{dT_x}{dx} = -l_x$$

and therefore,

$$\frac{d\overset{0}{e}_x}{dx} = -1 + \frac{T_x}{l_x^2}\left(-\frac{dl_x}{dx}\right) = \overset{0}{e}_x \mu_x - 1 \qquad (1)$$

Now the life expectancy is known to increase after age zero, to reach a maximum before age five and to decrease thereafter. Therefore, the only optimum value of $\overset{0}{e}_x$ is the maximum which is attained at say, $x = \hat{x}$, where

$$\overset{0}{e}_{\hat{x}}\mu_{\hat{x}} - 1 = 0$$

or

$$\overset{0}{e}_{\hat{x}} = 1/\hat{\mu}_x \qquad (2)$$

3. Minimum value of μ_x

For $\overset{0}{e}_x$ to be maximum at $x = \hat{x}$,

$$\frac{d^2\overset{0}{e}_x}{dx^2} = \mu_x\left(\overset{0}{e}_x\mu_x - 1\right) + \overset{0}{e}_x\frac{d}{dx}\mu_x \qquad (3)$$

must be negative at $x = \hat{x}$. Because of (2),

$$\left[\frac{d^2\overset{0}{e}_x}{dx^2}\right]_{x=\hat{x}} = \overset{0}{e}_{\hat{x}}\left[\frac{d}{dx}\mu_x\right]_{x=\hat{x}} \qquad (4)$$

which can be negative when

$$\left[\frac{d}{dx}\mu_x\right]_{x=\hat{x}} \text{ is negative.}$$

Again, μ_x is known to decline at age zero and to attain a minimum at, say $x = \tilde{x}$. Therefore, $d\mu_x/dx$ is negative in the age interval o to $\tilde{x}$. Clearly then, the maximum value of $\overset{0}{e}_x$ is obtained before μ_x reaches its minimum.

Thus, $\hat{x} < \tilde{x}$. (5)

4. Maximum value of the derivative of l_x

The minimum value of $\mu_x = -(dl_x/dx)/l_x$ is reached when

$$\frac{d\mu_x}{dx} = \frac{1}{l_x^2}\left(\frac{dl_x}{dx}\right)^2 - \frac{1}{l_x}\left(\frac{d^2l_x}{dx^2}\right) \qquad (6)$$

$$= \mu_x^2 - \frac{1}{l_x}\left(\frac{d^2l_x}{dx^2}\right) \qquad (7)$$

is zero at $x = \tilde{x}$, so that

$$l_{\tilde{x}}\mu_{\tilde{x}}^2 = \left[\frac{d^2l_x}{dx^2}\right]_{x=\tilde{x}} \qquad (8)$$

Now, the derivative of l_x is uniformly negative, and because of large force of mortality at age zero, assumes a maximum value at, say $x = \bar{x}$. Correspondingly, the second derivative of l_x is positive at age zero, decreases thereafter till it becomes zero at $x = \bar{x}$.

Since the right hand side of (8) is positive, it is clear that the minimum value of the force of mortality is reached before the derivative of l_x assumes its maximum or its smallest negative value. In other words,

$$\tilde{x} < \bar{x} \tag{9}$$

and combining (9) with (5), the following inequality relationship,

$$\hat{x} < \tilde{x} < \bar{x} \tag{10}$$

is established.

5. Estimation of $\bar{x}$, $\tilde{x}$ and $\hat{x}$

The exact age at which the three optimum conditions are met cannot be obtained since none of the life table functions l_x, μ_x, or $\overset{0}{e}_x$ can be expressed in terms of easily differentiable mathematical functions. Approximate solutions were however, obtained earlier (Mitra, ibid.) for ages corresponding to maximum life expectancies for different life tables. The method used was to find the point of intersection of two freehand curves obtained by plotting the values of $\overset{0}{e}_x$ and $1/\mu_x$ at early childhood ages. The values of μ_x can be determined from probabilities of dying ${}_nq_x$ in the age interval x to x+n, because of the well-known relationship

$$1-{}_nq_x = e^{-\int_0^n \mu_{x+t}\,dt} \tag{11}$$

The exponent $\int_0^n \mu_{x+t}\,dt$ can be approximated by $n\mu_{x+\frac{n}{2}}$ $(n<5)$

for the age interval 1 to 20 where the force of mortality is relatively small and finally, disregarding squares and higher powers of μ_x, (11) can be reduced to

$${}_nq_x = n\mu_{x+\frac{n}{2}} \tag{12}$$

Accordingly, the age, say x', at which ${}_nq_x$ is minimum, is approximately related to $\tilde{x}$ by the following equation,

$$\tilde{x} = x' + n/2 \tag{13}$$

The method of finite differences can then be used to determine x' for which ${}_nq_x$ values for n=5 and x=5, 10 and 15 are sufficient, since x' can generally be located around age ten. Using the conventional notation Δ for successive differences, so that

$$\Delta({}_nq_x) = {}_nq_{x+n} - {}_nq_x$$

and

$$\Delta^2({}_nq_x) = \Delta({}_nq_{x+n}) - \Delta({}_nq_x)$$

etc., the solution for $\tilde{x}$ is given by

$$\tilde{x} = 10 - \frac{5\Delta({}_5q_5)}{\Delta^2({}_5q_5)} \tag{14}$$

disregarding differences of third and higher orders.*

6. Applications

The formulas developed in the preceding sections for the estimation of $\hat{x}$, $\tilde{x}$ and $\bar{x}$, and their inequality relationships were tried on a few model life tables (United Nations, 1958) for males, selected to cover a wide range of expectations of life at birth. The results are shown in Table 1.

---

*According to Newton's forward difference formula, the function u_x, when u_0, u_1, u_2,...are known can be expressed as $u_x = u_0 + x\Delta u_0 + \frac{x(x-1)}{1.2}\Delta^2 u_0 + \ldots\ldots$ The function u_x can then be differentiated to determine optimum values, point of inflection, etc.

Table 1. Optimum values and respective ages of a few life table functions

Model life Table number	Expectation of life at birth $\overset{0}{e}_o$	Age of maximum life expectancy $\hat{x}$	Maximum life expectancy $\overset{0}{e}_{\hat{x}}$	Age of minimum force of mortality $\tilde{x}$	Age of maximum dl_x/dx $\overline{x}$
(1)	(2)	(3)	(4)	(5)	(6)
10	24.8	4.5	38.1	13.2	13.7
25	31.9	3.6	43.8	12.8	13.1
40	39.2	2.8	49.5	12.6	12.8
55	46.4	2.6	55.6	12.3	12.5
70	53.6	2.0	60.5	12.1	12.2
85	61.5	1.7	65.0	11.7	11.8
100	68.5	0.9	69.9	11.1	11.2

The sharp increase in life expectancy from age 0 to $\hat{x}$ is worth noting. So is the declining trend of $\hat{x}$ with increase in $\overset{0}{e}_o$. For life tables with lower life expectancies, the age of maximum value of $\overset{0}{e}_x$ is larger and this is due to high early childhood mortality in addition to high infant mortality. Both $\tilde{x}$ and $\overline{x}$ are large compared to $\hat{x}$, and their variations over the wide range of life tables are systematic and small. Like $\hat{x}$, $\tilde{x}$ and $\overline{x}$ are also inversely related with life expectancy, and cover a range of 13.2 to 11.1 and 13.7 to 11.2 respectively over the range of $\overset{0}{e}_o$ used in this analysis.

As noted in (9) $\tilde{x} < \overline{x}$, but the difference is rather small in most cases.

Summary

The characteristics of life table functions are quite well known but investigations of some of the crucial values of these functions and their interrelationships have not yet been carried out. The importance of such an inquiry need not be over-emphasized in view of the fact that any exercise on graduation of life table functions must take into account the order and spacing of these parametric values. It is known that the function l_x has a steep negative slope at age zero and the steepness continues to diminish till it becomes smallest at some age and begins to increase thereafter. Similarly, the force of mortality μ_x assumes a minimum value at some age and life expectancy $\overset{0}{e}_x$ attains a maximum value after age zero. These three ages have been found to form a sequence with the age corresponding to the maximum life expectancy as the lowest of the series. Methods have been outlined to estimate these values for a given life table and all three are found to be inversely related with expectation of life at birth.

References

Freeman, Harry, Mathematics for Actuarial Students, Vol. 2, Cambridge, 1949.

Mitra, S., "Graduation of Life Table Functions" American Statistical Association, Proceedings of the Social Statistics Section, 1971.

United Nations, Methods for Population Projections by Sex and Age, ST/SOA/Series A/Population Studies, No. 25, 1958.

SAMPLING 17-YEAR-OLDS NOT ENROLLED IN SCHOOL

R. P. Moore and B. L. Jones, Research Triangle Institute

INTRODUCTION

The purpose of the National Assessment of Educational Progress (A project of the Education Commission of the States) is to provide the public with data on the educational attainments of important groups in the population of the United States. To provide the basic data each year, probability samples of 9-, 13-, and 17-year-olds are assessed in elementary and secondary schools across the United States. Samples of young adults 26 to 35 years of age and 17-year-olds not enrolled in school are assessed in their homes. This paper describes the approach used to sample the out-of-school 17-year-old population in Year 02 of National Assessment and includes a limited description of the modifications made in Year 03.

THE BASIC PROBLEM

One of the National Assessment populations is defined as individuals who are 17 years of age (16-1/2 to 17-1/2) on April 1 of the assessment year. For example, the 17-year-old population for the Year 02 assessment was defined as individuals born between October 1, 1953, and September 30, 1954 (so 16-1/2 to 17-1/2 years old on April 1, 1971). A sample of 17-year-olds enrolled in school were surveyed in the Year 02 in-school assessment conducted during March, April, and May of 1971. A substantial number, perhaps 11 percent, of the 17-year-old population are not enrolled in elementary and secondary schools on the April first when they are 16-1/2 to 17-1/2 years of age (see Table 1).

TABLE 1 - PERSONS 16-1/2 TO 17-1/2 YEARS OF AGE ON APRIL 1, 1970, BY SCHOOL ENROLLMENT STATUS*

Age in years	Total number (000)	Enrolled grades K through 12		Not enrolled K through 12	
		Number (000)	Percent of total	Number (000)	Percent of total
16.50-16.75	1,038	951	91.6%	87	8.4%
16.75-17.00	923	834	90.4%	89	9.6%
17.00-17.25	946	824	87.1%	122	12.9%
17.25-17.50	949	810	85.4%	139	14.6%
16.50-17.50	3,856	3,419	88.7%	437	11.4%

*Source - Public Use Sample, 1 in 1,000 from 1970 U.S. Census.

The National Assessment program requires large numbers of respondents since several different instruments are used for each age group--usually about 12 different instruments per age group with a target sample size of 2,000 to 2,500 responses per instrument. The instruments must be administered by a trained interviewer in person, thus adding considerably to the expense and ruling out mailed inquiries. A multistage area sample of household residents, used to sample the 26- to 35-year-olds, was available for use in surveying the out-of-school 17-year-olds, but the anticipated number of out-of-school 17-year-old respondents from this sample was only about 100 per year, while 600 to 800 respondents were needed each year. Increasing the size of the area sample this much was considered too expensive since it was necessary to screen approximately 100 households in order to locate one eligible out-of-school 17-year-old.

YEAR 02 PILOT STUDY

Introduction

Several potentially useful list frames were considered in Year 02 of National Assessment (1970-71 school year). Table 2 shows the frames considered and several relevant characteristics of each. The secondary school records frame, consisting of dropouts reported by secondary schools during the past three school years, was the most promising list frame. A pilot survey was undertaken to gain experience in the operational and analytical problems involved in using the secondary school records frame.

The Neighborhood Youth Corps and Job Corps frames, though limited in universe coverage, were also included in the pilot study since these frames were readily available, assessment costs were expected to be relatively low due to the possibility of group testing, and since the Job Corps frame consisted mainly of individuals in group quarters (not covered by the area frame of households).

TABLE 2 - SUMMARY OF ALTERNATIVE SAMPLING FRAMES INVESTIGATED

Frame	Expected universe coverage	Relative sampling and assessment costs	Group sessions feasible	Difficulty in constructing frame	Covered by area (household) frame?	Anticipated cooperation (of agencies)
Area	95%	High	No	None	Yes	---
Secondary School Records	70 to 80%	Low	No	Some	Most	Some problems
Colleges	2 to 3%	Medium	No	Considerable	Partial	Some problems
Military	3 to 4%	Very low	Yes	Some	Partial	Very poor
Neighborhood Youth Corps	5 to 6%	Very low	Yes	None	Yes	Good
Job Corps	1 to 2%	Very low	Yes	None	No	Very good
Employment Security Comm. (active files)	1 to 5%	Low	No	Considerable	Most	Some problems

The area frame was also sampled in the Year 02 pilot study, since it was the most complete sampling frame available and would be used to survey the 26- to 35-year-olds anyway. The sample from the area frame was limited to the number of households to be surveyed for young adults 26 to 35 years of age.

In order to increase the out-of-school 17-year-old sample size from the area frame, individuals 17-1/2 to 18-1/2 years old who were not enrolled in school when they were 16-1/2 to 17-1/2 were also regarded as eligible. It was assumed that these individuals were representative of the population of interest. This allowed a larger sample size to be obtained from the area sample.

Estimation Procedure and Assumptions

Multiple frame sampling [1] describes the situation where several frames are sampled independently in the course of a single survey. Multiple frames are often used when either there is no single complete frame or an "expensive" complete frame is used jointly with one or more incomplete but "cheap" frames. Two assumptions are required: (1) each population member belongs to at least one of the frames used, and (2) the association (or lack of association) of each sample individual with each of the frames surveyed can be determined.

Each sample individual is classified as being a member of one of a number of domains. Domain totals are estimated using the sample data for each domain of each frame separately. Since the frames overlap, more than one estimate will be produced for some domains. All estimates for a domain are weighted together to obtain one overall estimate for each domain. The weighted domain total estimates are then simply summed to estimate population totals.

The model which defines the domains of interest for this study is shown in Figure 1. Sampling frames are identified by capital letters while domains are denoted by lower case letters. The following assumptions were made in constructing the domain model shown in Figure 1.

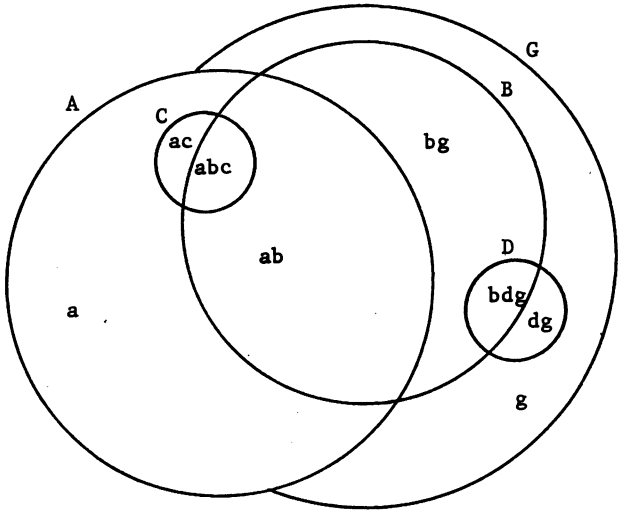
(a) Frames A and G do not overlap and their union is complete. That is, every eligible out-of-school 17-year-old belongs to the household population or the group quarters population, but not both.

(b) Frame C is a subset of frame A. That is, all Neighborhood Youth Corps members are in the household population.

(c) Frame D is a subset of frame G. That is, all Job Corps enrollees are in the group quarters population.

Given these assumptions, the four sampled frames intersect to define the seven domains shown in Figure 1. In order to completely specify the population, frame G is also shown in the figure. Estimates for domain g are not available from the survey data since frame G was not sampled.

Domains defined by several additional frames which were not sampled were also estimated and analyzed [2], but are not discussed in this paper. The domain estimates by frames and the weighted estimates are presented in subsequent sections of this paper.



Frames: A, Area frame; B, School dropout lists; C, Neighborhood Youth Corps; D, Job Corps Centers; G, Group quarters (not sampled).

FIGURE 1 - THE DOMAIN MODEL FOR THE YEAR 02 PILOT STUDY

Secondary School Records Sample

A sample of schools was selected by subsampling the in-school sample [3] used for the assessment of 17-year-olds in Year 02. This was a multistage probability sample of 116 first-stage units (counties or groups of counties) with clusters of schools as second-stage units, and schools having one or more of grades 9, 10, 11, or 12 as third-stage units. First, 58 of the 116 sample PSUs were selected for the pilot study sample. Next, 173 of the 347 sample schools for the assessment of in-school 17-year-olds in the 58 selected PSUs were selected.

Lists of dropouts during three school years (1968-69, 1969-70, 1970-71) were requested and cooperation was obtained from 147 or 85.0 percent of the 173 sample schools (see Table 3). After screening the lists to eliminate nonpopulation members who could be identified by their birthdates or school withdrawal dates, a sample of 957 potential eligibles was selected.

Field interviewers were assigned to locate the selected dropouts, ascertain their eligibility status, and administer assessment packages to those who were eligible (members of the target population of out-of-school 17-year-olds). The eligibility status was determined for 701 or 73.2 percent of the total sample, as shown in Table 4. The "not eligible" category included 43 persons whose birthdates did not meet the population definition and 182 who were enrolled in school during the reference month. This is not surprising since

TABLE 3 - SCHOOL RESPONSE TO REQUEST FOR LISTS OF DROPOUTS, YEAR 02

Category	Number of schools	Percent
Provided dropout lists	147	85.0
Refused to participate	26	15.0
Total	173	100.0

TABLE 4 - RESPONSE RATES FOR SAMPLE OF SCHOOL DROPOUTS, YEAR 02

Category	Number	Percent
Eligibility status:		
Eligible	476	49.7
Not eligible	225	23.5
Not determined	256	26.8
Total sample	957	100.0
Response status:		
Respondent	345	72.5
Refused	89	18.7
Not located	42	8.8
Total eligibles	476	100.0

some of the schools providing dropout lists did not supply the birthdate and withdrawal date information. Those whose eligibility was not determined were primarily cases where neither the dropout nor any other knowledgeable family member could be located.

The 345 respondents (Table 4) completed a total of 1,317 assessment packages or an average of 3.82 per respondent. (A total of 12 different packages were used to assess 17-year-olds in Year 02. Each out-of-school respondent was offered an incentive payment of five dollars per package if he agreed to complete two, three, or four packages. Each package required approximately one hour of the respondent's time.)

A set of domain classification questions such as "Do you belong to the Neighborhood Youth Corps?" was asked of each sample person. Based on the responses, each respondent was classified as a member of either domain ab, abc, bg, or bdg. A number of problems were encountered in making the classification due to missing data and respondents misinterpreting the domain classification questions [2].

Table 5 shows the domain estimates obtained by weighting the counts of eligibles by domain. Weights were computed as the inverses of the overall selection probabilities, with appropriate adjustments for nonresponse at all levels. Table 5 also shows the expanded estimates for 17-year-olds and for 18-year-olds separately.

Neighborhood Youth Corps and Job Corps Samples

The Year 02 pilot study included a small amount of work in sampling Neighborhood Youth Corps (NYC) and Job Corps (JC) enrollees. Samples of five NYC centers and five JC centers were selected. Within each center, the plan was to assess four enrollees individually and 24 enrollees in two groups of twelve per group. Each sampled enrollee was asked to complete two assessment packages; an incentive payment of ten dollars per respondent was offered.

List frames for the NYC (out-of-school component) and the JC centers were obtained from the agencies' headquarters in Washington, D.C. The NYC centers were stratified by seven size-of-center categories and four geographic regions. The JC centers were stratified by five type-of-center descriptions and four regions. A controlled selection procedure was used to select a sample of five NYC centers and five JC centers with probabilities proportional to size. The

TABLE 5 - DOMAIN TOTALS ESTIMATED FROM SCHOOL DROPOUT LISTS SAMPLE, BY AGE GROUP, YEAR 02

Domain	Estimated total 16-1/2 to 17-1/2 (born 10/53 to 9/54)	Estimated total 17-1/2 to 18-1/2 (born 10/52 to 9/53)	Estimated total eligibles
ab	139,935	142,925	282,863
abc	5,945	5,358	11,300
bg	3,648	9,787	13,435
bdg	1,744	0	1,744
Total	151,272	158,070	309,342

size measures were authorized enrollments (NYC) and center capacities (JC).

One of the selected NYC centers could not be located and it was later determined that the center had not operated since 1968. The four remaining NYC centers were all in operation; one of the four refused to participate in the assessment. All five sampled JC centers were operating and agreed to participate.

Records at the NYC and JC centers were examined to determine a list of all enrollees in the centers who were eligible for the study (belonged to the survey population by birthdate and school withdrawal date). A total of 318 assessment packages were completed by 159 sample enrollees in the eight cooperating NYC and JC centers. There were difficulties in that some enrollees did not keep appointments, especially in the NYC centers where enrollees were working at scattered locations rather than at the centers.

The domain estimates computed from the NYC and JC sample data are shown in Table 6. The numbers of respondents are also shown. Estimates are the sums of weights by domain and age group; weights were computed as the inverses of selection probabilities, adjusted for nonresponse.

The Area Sample

The National Assessment Year 02 out-of-school sample [3] was also used to survey out-of-school 17-year-olds. The multistage area probability

TABLE 6 - DOMAIN TOTAL ESTIMATES AND NUMBERS OF RESPONDENTS FOR NEIGHBORHOOD YOUTH CORPS AND JOB CORPS SAMPLES, BY AGE GROUP*

Sample and domain	Estimated total 16-1/2 to 17-1/2 (born 10/53 to 9/54)	Estimated total 17-1/2 to 18-1/2 (born 10/52 to 9/53)	Estimated total eligibles
Neighborhood Youth Corps:			
ac	(7) 1,266	(9) 2,391	(16) 3,657
abc	(14) 3,324	(14) 2,966	(28) 6,291
Total	(21) 4,590	(23) 5,358	(44) 9,948
Job Corps:			
dg	(13) 718	(8) 296	(94) 1,014
bdg	(54) 2,280	(40) 1,165	(94) 3,446
Total	(67) 2,998	(48) 1,461	(115) 4,460

*Estimates may not add to totals because of rounding.

TABLE 7 - NATIONAL ASSESSMENT YEAR 02
OUT-OF-SCHOOL SURVEY RESPONSE EXPERIENCE
17-YEAR-OLDS

Item	Total	Percent of total	Average per segment
Occupied housing units	8,203	100.0	15.78
Housing units screened	8,131	99.1	15.64
Reason for nonscreening:			
Not at home	41	0.5	.079
Incompetent	3	0.1	.006
Refused	28	0.3	.054
Eligible 17's			
out-of-school	86	100.0	.165
Package Respondents	83	96.5	.160
Reason for nonresponse:			
Not at home	1	1.2	.002
Refused	2	2.3	.004

design consisted of 52 first-stage units (counties or groups of counties) and 520 second-stage units (clusters of housing units). Primary units were stratified by region, socioeconomic status, and size of community. The low socioeconomic stratum was sampled at twice the rate of the remaining stratum. Third-stage units were the individual housing units determined by field listing. All eligible out-of-school 17-year-olds living in all sample housing units were "in the sample."

A total of 8,131 of the 8,203 sample housing units in the 520 second-stage sampling units were screened for out-of-school 17-year-olds (see Table 7). Of the 86 eligibles identified, 83 cooperated and completed a total of 325 assessment packages, an average of 3.92 per respondent.

Domain estimates were computed by summing the weights of sample eligibles by domains. Weights were calculated as the inverses of selection probabilities, adjusted for nonresponse. Table 8 shows the domain total estimates by domain, and by age group. The numbers of respondents are shown in parentheses. Adding the estimates for domains ab and abc indicates that 428,742 of the estimated 551,016 in the household population, or 78 percent, were associated with the secondary school records (dropout list) frame.

Combined Domain Total Estimates

The domain total estimates shown in the previous three sections were computed using the sample data from each domain of each sampled frame separately. The eight domains defined by the model (Figure 1) are shown in Table 9 along with the domain total estimates. Also shown are combined domain total estimates obtained by weighting together the domain estimates for overlapping domains. The weights used to combine two or more estimates for a particular domain should meet two conditions: (1) the weights should be determined independently of the survey estimates, and (2) the sum of the weights applied to the various estimates for a domain must sum to one. The weights used to obtain the overall domain estimates shown in Table 9 were computed proportionate to the average first-stage sampling rates.

The weighted domain totals should be the best domain size estimates available; the sum of the weighted estimates is an estimate of the Year 02 survey population, defined by the union of frames A, B, C, and D (Figure 1). The estimates shown are for both 17-and 18-year-olds and are not adjusted to an estimate for a one-year age group.

Coverage of Population

In order to estimate the population coverage afforded by various frames and unions of frames, it is necessary to estimate the universe size. The area or household frame covers approximately 93.4 percent of the population, based on data from the 1970 Census Public Use Sample (see Table 10). An estimate of the total population size may be obtained by dividing .934 into the sum of the weighted domain total estimates for the domains

TABLE 8 - DOMAIN TOTAL ESTIMATES AND NUMBERS OF RESPONDENTS FOR THE AREA SAMPLE, BY AGE GROUP*, YEAR 02

Domain	Estimated total 16-1/2 to 17-1/2 (born 10/53 to 9/54)	Estimated total 17-1/2 to 18-1/2 (born 10/52 to 9/53)	Estimated total eligibles
a	(7) 39,411	(14) 77,461	(21) 116,872
ab	(29) 178,880	(30) 236,025	(59) 414,904
ac	(0) 0	(1) 5,403	(1) 5,403
abc	(0) 0	(2) 13,837	(2) 13,837
Total	(36) 218,291	(47) 332,726	(83) 551,016

*Estimates may not add to totals due to rounding.

TABLE 9 - DOMAIN TOTAL ESTIMATES, BY FRAMES, AND WEIGHTED DOMAIN TOTAL ESTIMATES, YEAR 02

Domain	Estimates computed from sample data from frame				Weighted domain totals
	A	B	C	D	
a	116,872	---	---	---	116,872
ab	414,904	282,863	---	---	328,945
ac	5,403	---	3,657	---	4,968
abc	13,837	11,300	6,291	---	11,573
bg	---	13,435	---	---	13,435
bdg	---	1,744	---	3,446	2,828
dg	---	---	---	1,014	1,014
g*	---	---	---	---	---
Total	551,016	309,342	9,948	4,460	479,635

*A total for domain g cannot be estimated from the survey data.

TABLE 10 - PERSONS 16-1/2 TO 17-1/2 YEARS OF AGE AND NOT ENROLLED IN GRADES K THROUGH 12 ON APRIL 1, 1970, BY TYPE OF RESIDENCE*

Age in years	Total number (000)	Household population Number (000)	Household population Percent of total	Group quarters population Number (000)	Group quarters population Percent of total
16.50-16.75	87	84	96.6%	3	3.4%
16.75-17.00	89	86	96.6%	3	3.4%
17.00-17.25	122	112	91.8%	10	8.2%
17.25-17.50	139	126	90.6%	13	9.4%
16.50-17.50	437	408	93.4%	29	6.6%

*Source - Public Use Sample, 1 in 1,000 from 1970 U.S. Census.

TABLE 11 - ESTIMATED FRAME SIZES AND POPULATION COVERAGE FOR VARIOUS FRAMES AND UNIONS OF FRAMES, YEAR 02

Frame	Estimated[a] frame size (number)	Estimated population coverage (percent)
A	462,359	93.4
B	356,782	72.1
C	16,541	3.3
D	3,842	0.8
G[b]	17,277	3.5
BUC	361,750	73.1
BUD	357,796	72.3
BUCUD	362,764	73.3
AUB	478,622	96.7
AUBUC	478,622	96.7
AUBUD	479,636	96.9
AUBUCUD	479,636	96.9
Total[c]	495,031	100.0

[a]Estimates for frames and unions of frames obtained by summing the appropriate weighted domain totals.

[b]Incomplete estimate since domain g cannot be estimated.

[c]This estimate is based on the assumption that frame A is 93.4 percent complete.

included in frame A (462,359 ÷ .934 = 495,031). Comparisons of frame coverage were computed and are shown in Table 11.

Frame B (school dropout lists) is estimated to cover about 72 percent of the population of interest. Frames C (Neighborhood Youth Corps) and D (Job Corps) cover a small part of the population. Attempting to maximize the part of the population covered by the "cheap" frames (B, C, and D) did not appear very promising since BUCUD was estimated to cover only 1.2 percent more than B alone. Nearly 97 percent coverage can be obtained by sampling frames A (area frame) and B (school dropout lists); additional coverage obtained by also sampling from frames C and D would be negligible.

Level of the Estimated Totals

The estimated total for the out-of-school 17-year-old population (495,031 for a two-year age group) shown in Table 11 looks low compared with Census Public Use Sample Data (437,000 for a one-year age group from Table 10) and Current Population Survey estimates. A small part of the difference occurred since mentally and physically handicapped persons were not counted as eligibles and the withdrawal date used in defining the survey population was adjusted by about three months to accommodate the data collection schedule. The domain ab total estimated from the frame B sample looks low (see Table 9). A possible explanation is that some schools do not have sufficiently adequate records to prepare a complete list of all dropouts during the past three school years.

Summary of Results

The pilot study results were interpreted to indicate that the secondary school records frame and the area frame should be used in combination for subsequent years of National Assessment. The alternative of increasing the size of the area sample by several times was not considered practical.

It was suggested that, although the union of the area and school dropout frames covered an estimated 97 percent of the population, the coverage might be increased slightly by surveying lists of early high school graduates obtained from a sample of high schools. An increase in the coverage of the school dropout frame was also hypothesized if one were to obtain dropout lists from schools with grades 7 or 8 in addition to grades 9, 10, 11, or 12.

The response rates attained in the pilot study were regarded as satisfactory, but it was hoped that better selection and training of interviewers would result in fewer sample dropouts in the "not located" category in Year 03.

THE YEAR 03 SURVEY

This section briefly describes the modifications in the overall survey design for Year 03 of National Assessment. A more complete document describing the Year 03 assessment is available [4].

Four frames were sampled in Year 03 (1971-72 school year) to survey out-of-school 17-year-olds:

- Frame A - area frame,
- Frame B - secondary school records frame of dropouts from schools with grades 9, 10, 11, or 12,
- Frame H - early high school graduates reported by frame B schools, and
- Frame J - secondary school records frame of dropouts reported by schools with grades 7 or 8 but none of the grades 9, 10, 11, and 12.

Frames A and B were defined in the same way as in Year 02. Frame H consisted of persons reported to have graduated and left school before the April first date when they were 16-1/2 to 17-1/2 years of age. Since many frame H members were enrolled in colleges and living in group quarters, they were not covered by frames A and B. Frame J was added to determine whether or not the population coverage by the "cheap" frames might be increased by sampling dropouts from schools with any of the grades 7 through 12 rather than only those with any of the grades 9 through 12. Frame J was kept separate from frame B in order to evaluate this difference in coverage.

The sample designs for the frame A and frame B surveys were similar to the Year 02 designs. The sample sizes were approximately double those for Year 02. The frame A sample consisted of 104 primary units and 1,040 secondary units and over 17,000 households. The school samples for frames B, H, and J were selected from the 116 Year 03 in-school sample primary units. The sample sizes in terms of numbers of schools selected are shown in Table 12, which also shows response rates for determining eligibility status of sample individuals and cooperation rates for those determined to be eligible for the survey. The response rates were similar to those for Year 02. The response rates for the frame A sample in Year 03 are shown in Table 13.

The domain total estimates by frames and the overall domain total estimates for Year 03 are shown in Table 14. The methods used to compute

TABLE 12 - SCHOOL AND INDIVIDUAL RESPONSE RATES, YEAR 03 LIST SAMPLES

Item	Frame B		Frame H		Frame J	
	No.	Percent	No.	Percent	No.	Percent
School response:						
Provided lists	362	85.0	103	79.2	199	93.0
Refused	64	15.0	27	20.8	15	7.0
Total	426	100.0	130	100.0	214	100.0
Eligibility status:						
Eligible	1,024	51.1	55	57.9	64	46.0
Not eligible	500	25.0	17	17.9	53	38.1
Not determined	479	23.9	23	24.2	22	15.8
Total	2,003	100.0	95	100.0	139	100.0
Response status:						
Respondent	732	71.5	39	70.9	45	70.3
Refused	188	18.4	11	20.0	11	17.2
Not located	104	10.2	5	9.1	8	12.5
Total	1,024	100.0	55	100.0	64	100.0

TABLE 13 - YEAR 03 FRAME A SURVEY RESPONSE FOR OUT-OF-SCHOOL 17-YEAR-OLDS

Item	Total	Percent of total	Avg. per segment
Occupied housing units	17,184	100.0	18.36
Housing units screened	17,126	99.7	18.30
Reason for nonscreening:			
Not at home	17	0.1	.02
Refused	41	0.2	.04
Eligible 17's out-of-school	158	100.0	.169
Package respondents	139	88.0	.149
Reason for nonresponse:			
Not at home	5	3.2	.005
Refused	14	8.9	.015

the estimates were the same as have been described for Year 02. The Year 03 domain model assumed that frames B, H, and J were nonoverlapping and that each of those three frames overlapped with frames A and G.

Table 15 shows the estimated frame sizes and the estimated population coverage for frames and unions of frames. The estimated coverage by frames A and B was about the same as in Year 02. Sampling from frame H in addition to frames A and B increased the estimated coverage from 96.2 percent to 96.5 percent, but the coverage by the "cheap" frames could be increased from 73.5 percent (frame B) to 78.7 percent (frames B and H). For this reason, it appeared that sampling frame H was worthwhile. Also the dropout and early graduate lists could be obtained from the same sample schools, so the additional cost of adding frame H was small. Frame J added little to the population coverage and the use of frame J was not continued beyond Year 03.

The level of the estimates was still low, as in Year 02, and appeared to indicate that schools were not able to supply complete lists of dropouts during the previous three school years.

TABLE 14 - DOMAIN TOTAL ESTIMATES, BY FRAMES, AND WEIGHTED DOMAIN TOTAL ESTIMATES, YEAR 03

	Domain estimates computed from sample data from frame:				
Domain	A (area)	B (dropouts from 9,10, 11,12 grade schools)	H (early high school grads)	J (dropouts from 7,8 grade schools)	Weighted domain totals
a	77,708	—	--	--	77,708
bg	--	13,267	--	--	13,267
gh	--	--	1,699	--	1,699
gj	--	--	--	575	575
ab	475,238	263,426	--	--	339,149
ah	36,552	--	16,377	--	23,590
aj	7,182	--	--	7,909	7,649
g*	--	--	--	--	--
Total	596,680	276,693	18,076	8,484	463,637

*A domain g total cannot be estimated from the survey data.

REFERENCES

[1] H. O. Hartley, "Multiple Frame Surveys," Proceedings of the Social Statistics Section of the American Statistical Association meeting, Minneapolis, Minn., 1962.

[2] R. P. Moore and B. L. Jones, "Study of Alternative Sampling Frames for Out-of-School 17-Year-Olds," Technical Report No. 1 on RTI Project 25U-688-1, Research Triangle Institute, December 1971.

[3] J. R. Chromy, R. P. Moore, and W. T. Rogers, "The National Assessment Approach to Sampling--Year 02," National Assessment of Educational Progress (In press), Denver, Colorado, 1973.

[4] R. P. Moore and B. L. Jones, "Multiple Frame Sampling for Out-of-School Seventeen-Year-Olds In Year 03 of National Assessment," Technical Report No. 1 on RTI Project 25U-796-3, Research Triangle Institute, February 1973.

TABLE 15 - ESTIMATED FRAME SIZES AND POPULATION COVERAGE FOR VARIOUS FRAMES AND UNIONS OF FRAMES, YEAR 03

Frame	Estimated[a] frame size (number)	Estimated Population coverage (percent)
A	448,096	93.4
B	352,416	73.5
H	25,289	5.3
J	8,224	1.7
G[b]	15,541	3.2
BUH	377,705	78.7
BUJ	360,640	75.2
BUHUJ	385,929	80.4
AUB	461,363	96.2
AUBUH	463,062	96.5
AUBUHUJ	463,637	96.6
Total[c]	479,760	100.0

[a,b,c]See Table 11 footnotes.

ESTIMATING THE VARIANCES OF SAMPLE CORRELATIONS

Alice Orcutt Nakamura and Masao Nakamura, University of Alberta

I. INTRODUCTION

Let

$$p(r|r_1,r_2,\rho_1,\rho_2) = \frac{p(r,r_1,r_2|\rho_1,\rho_2)}{p(r_1,r_2|\rho_1,\rho_2)} \tag{1}$$

denote the distribution of the sample product moment correlation coefficient, r, conditional on the first-lag sample autocorrelations, r_1 and r_2, and the first-lag population autocorrelations ρ_1 and ρ_2, where the population cross-correlation is $\rho = 0$. The conventional tests of significance for r rely on

$$p(r|\rho_1,\rho_2) = \iint p(r,r_1,r_2|\rho_1,\rho_2)dr_1dr_2, \tag{2}$$

Moreover Bartlett [1] and McGregor [5,6] have established that

$$p(r|\rho_1,\rho_2) \cong p(r|\rho_1\ \rho_2). \tag{3}$$

and that

$$V(r|\rho_1,\rho_2) \cong V(r|\rho_1\ \rho_2) \tag{4}$$

$$\cong \frac{1+\rho_1\rho_2}{n(1-\rho_1\rho_2)} - \frac{2\rho_1\rho_2(1-\rho_1^n\rho_2^n)}{n^2(1-\rho_1\rho_2)^2} \tag{5}$$

$$\cong \frac{1+\rho_1\rho_2}{n(1-\rho_1\rho_2)} - \frac{2\ \rho_1\rho_2}{n^2(1-\rho_1\rho_2)^2} \tag{6}$$

where n denotes the number of items in each of the series correlated.[1]

In a recent paper, Nakamura, Nakamura and Orcutt [7] note that (1) is more informative than (2), and argue that tests of significance for r should be based on (1) rather than (2). Moreover Monte Carlo evidence is presented that

$$V(r|r_1,r_2,\rho_1,\rho_2) = V(r|r_1,r_2). \tag{7}$$

Lacking Monte Carlo tabulations of (7) for small samples, a researcher can attempt to estimate (7) by substituting r_1 and r_2 for ρ_1 and ρ_2 in equation (6) as suggested by Orcutt and James [8], he can use (6) directly, or he can use

$$V(r|\rho_1\rho_2 = 0) \cong \frac{1}{n-3}^2. \tag{8}$$

As $n \to \infty$ clearly $V(r|r_1,r_2,\rho_1,\rho_2) = V(r|r_1,r_2) \to V(r|\rho_1,\rho_2)$.

In this paper, sampling methods are used to compare the percentage errors made in estimating (7) for series of length 30 using these three approaches.

II. METHODOLOGY

Our generating relationships were of the form

$$X_t = \rho_1 X_{t-1} + u_t, \tag{9}$$

and

$$Y_t = \rho_2 Y_{t-1} + v_t, \tag{10}$$

where u and v were generated by two Chen random normal number generators.[3] 1,000 pairs of series of length 30 were generated and saved for each of the following pairs of values of ρ_1 and ρ_2: (-.9,.9),(-.7,.7),(-.3,.3),(-.3,-.3),(.3,.3), (-.7,-.7),(.7,.7),(-.9,-.9), and (.9,.9).[4]

For each series the autoregressive parameter was estimated using least squares regression.[5] Also we calculated the Pearson product-moment correlation coefficient for each pair of series. Each set of 1,000 sample correlation coefficients was then classified according to the values of the products of the sample autocorrelation coefficients, r_1r_2, of the pairs of series correlated. Intervals of 0.1 were used. Finally the observed variance of the sample correlations was calculated for each cell for each of our 9 sets of 1,000 correlations.

For each cell in our product classification for each of our 9 sets of correlations we next approximated the variance of the sample correlations in that cell using the modified version of formula (6):

$$\text{var } r \cong \frac{1 + \overline{r_1r_2}}{30(1-\overline{r_1r_2})} - \frac{2\ (\overline{r_1r_2})}{900(1-\overline{r_1r_2})}2 \tag{11}$$

where $\overline{r_1r_2}$ stands for the cell mean of the products of the sample autoregressive coefficients. Secondly we estimated the cell variances using formula (6) with n = 30. As a third alternative, we estimated the cell variances using formula (8). We will call the estimates obtained for each cell using formulas (11), (6) and (8), estimates 1,2 and 3 respectively.[7]

We now calculated the percentage errors made in approximating the observed cell variances of our sample correlations using each of these three estimation methods. The formula used to obtain these percentage errors was

$$\%\text{ error } i = \frac{(\text{estimate } i)-(\text{observed cell variance})}{(\text{observed cell variance})} \tag{12}$$

$$i = 1,\ 2,\ 3.$$

The percentage errors are shown in Table 1, where the top number in each cell corresponds to the percentage error made using estimate 1, the next

number to the percentage error made using estimate 2, and the third number to the percentage error made using estimate 3 for that cell. The cell frequencies - that is, the number of correlations in each cell - are shown in Table 2.

III. FINDINGS

Estimation method 1 results in smaller percentage errors in estimating our observed cell variances than either estimation methods 2 or 3 for 69% of our cells, and smaller percentage errors than method 2 for 80% of our cells. Looking only at those cells where the frequency, or number of correlations, is at least 30, and hence where the observed sample variances of the correlations in each cell can be regarded as a reasonable estimate of the population conditional variance for that cell, we see that method 1 results in smaller percentage errors than either methods 2 or 3 for 84% of these 49 cells.

Thus method 1 is seen to be a more efficient method of estimating $V(r|r_1,r_2)$ than either methods 2 or 3,[8] and is more operational than method 2 which requires knowledge of the population autoregressive parameters. Further experiments using (-.9,0),(-.7,0),(-.3,0),(0,0),(.3,0),(.7,0),(.9,0) for the values of ρ_1 and ρ_2 indicate that this result holds even when $\rho_1\rho_2 = 0$.

FOOTNOTES

1. This formalization of our problem was suggested to us by Professor Arthur S. Goldberger.
2. See Fisher [4], p. 191.
3. See Chen [2,3]. The initial values used for the starting integers were 748511649 and 147303541 for the u series and 180810529 and 536841077 for the v series. Satisfactory statistical properties are reported for random numbers generated using these initial numbers in Chen [3]. For both series the mean was 0 and the standard deviation was 25. We set $X_0 = Y_0 = 0$. The computer used was the IBM System/360 model 67 at the University of Alberta Computing Center.
4. To minimize the effect of the initial values used in generating u and v the first pair of series of length 30 generated for each pair of values of ρ_1 and ρ_2 was discarded. Also every other one of the subsequent pairs of series of length 30 generated was discarded.
5. Since in practice one would have no way of knowing the true value of the constant term, we estimated a constant term along with the autoregressive parameter.
6. Since
$$E(r|r_1r_2) = E(r) = 0$$
where r denotes the sample correlation coefficient [7],
$$\overline{r^2} = \frac{\Sigma r^2}{n} = \frac{\Sigma[r-E(r)]^2}{n}$$
is an unbiased estimate of the variance of r. This is the formula which we used in computing the cell variances.
7. In our abstract these three estimation methods are referred to in reverse order.
8. Stuart [9] presents a theoretical argument showing that given an estimator u of a parameter θ in a multiparameter distribution, one does not necessarily improve its efficiency by substituting true parameter values into u to replace estimators of them. For a discussion of estimating efficiency and the power of tests see Sundrum [10].

REFERENCES

[1] Bartlett, M.S., "Some Aspects of the Time Correlation Problem in Regard to Tests of Significance," Journal of the Royal Statistical Society, 98 (1935), 536-43.

[2] Chen, E.H., "A Random Normal Number Generator for 32-Bit-Word Computers," Journal of the American Statistical Association, 66 (1971), 400-3.

[3] _____, "Supplement to 'A Random Normal Number Generator for 32-Bit-Word Computers,'" Health Sciences Computing Facility, University of California, Los Angeles, California, 1971.

[4] Fisher, R.A., Statistical Methods for Research Workers, twelfth edition, Edinburgh: Oliver and Boyd, 1954.

[5] McGregor, J.R., "The Approximate Distribution of the Correlation between Two Stationary Linear Markov Series," Biometrika, 49, Nos. 3 and 4 (1962), 379-88.

[6] _____, and U.M. Bielenstein, "The Approximate Distribution of the Correlation between Two Stationary Linear Markov Series II," Biometrika, 52 (1965), 301-2.

[7] Nakamura, A.O., M. Nakamura, and G.H. Orcutt, "Testing for Relationships between Time Series," an unpublished paper.

[8] Orcutt, G.H., and S.T. James, "Testing the Significance of Correlation between Time Series," Biometrika, 35, Nos. 3 and 4 (1948), 397-413.

[9] Stuart, A., "A Paradox in Statistical Estimation," Biometrika, 42 (1955), 527-9.

[10] Sundrum, R.M., "On the Relation between Estimating Efficiency and the Power of Tests," Biometrika, 41 (1954) pp. 542-4.

1. PERCENTAGE ERRORS

Class intervals for $r_1 r_2$	Values of ρ_1 and ρ_2								
	(-.9,.9)	(-.7,.7)	(-.3,.3)	(-.3,-.3)	(.3,.3)	(-.7,-.7)	(.7,.7)	(-.9,-.9)	(.9,.9)
	-114								
-1.1 to -1.0	7								
	875								
	-36								
-1.0 to -.9	44								
	1221								
	15	444							
-.9 to -.8	35	1638							
	1134	5309							
	-6	83							
-.8 to -.7	-29	278							
	552	1076							
	6	19							
-.7 to -.6	-43	81							
	419	462							
	-3	7							
-.6 to -.5	-61	22							
	259	281							
	1	-4	22						
-.5 to -.4	-68	-14	140						
	190	169	217						
	-12	2	2						
-.4 to -.3	-78	-26	64						
	100	131	117						
	4	-2	2						
-.3 to -.2	-78	-42	37						
	97	80	82						
	-9	13	-6	25	-14				
-.2 to -.1	-85	-45	3	91	33				
	39	71	36	78	24				
	298	-18	-4	-3	-4		18		
-.1 to 0	-45	-66	-12	21	20		252		
	401	5	16	13	12		40		
		292	-3	0	-14	-40	-10		-29
0 to .1		35	-22	7	-5	45	119		405
		320	3	0	-12	42	-13		-30
			664	2	12	-12	-13	9	-32
.1 to .2			390	-8	2	81	80	570	292
			548	-14	-5	-28	-28	-7	-46
				-17	13	31	-4	-15	-22
.2 to .3				-39	-14	120	62	293	282
				-43	-20	-13	-36	-46	-47

Class intervals for $r_1 r_2$	Values of ρ_1 and ρ_2					
	(-.9,.9)	(-.7,.7)	(-.3,.3)	(-.3,-.3)	(.3,.3)	(-.7,-.7)	(.7,.7)	(-.9,-.9)	(.9,.9)
.3 to .4	-5	220	10	-9	-8	32
	-41	81	51	27	262	422
	-45	69	-40	-50	-50	-28
.4 to .5	10		7	-14	38	-9
	-47		19	-6	327	184
	-50		-53	-63	-41	-61
.5 to .6			3	-10	5	-12
			-11	-22	155	115
			-64	-69	-65	-70
.6 to .7			4	60	24	-15
			-31	5	123	57
			-73	-58	-69	-78
.7 to .8			64	51	10	-9
			-19	-25	40	20
			-68	-70	-80	-83
.8 to .9			361		16	-4
			68		-2	-15
			-33		-86	-88
.9 to 1.0					43	55
					-22	-16
					-89	-88
1.0 to 1.1						

2. CELL FREQUENCIES

Class intervals for $r_1 r_2$	Values of $\rho_1\rho_2$								
	(-.9,.9)	(-.7,.7)	(-.3,.3)	(-.3,-.3)	(.3,.3)	(-.7,-.7)	(.7,.7)	(-.9,-.9)	(.9,.9)
-1.1 to -1.0	1								
-1.0 to -.9	14								
-.9 to -.8	118	1							
-.8 to -.7	251	9							
-.7 to -.6	280	64							
-.6 to -.5	177	139							
-.5 to -.4	94	279	1						
-.4 to -.3	45	270	9						
-.3 to -.2	12	154	55						
-.2 to -.1	6	66	246	2	4				
-.1 to 0	2	16	544	99	186		4		
0 to .1		2	142	485	575	5	35		3
.1 to .2			3	307	199	36	117	1	13
.2 to .3				93	34	108	212	7	40
.3 to .4				13	2	205	288	21	84
.4 to .5				1		289	214	33	153
.5 to .6						226	107	91	249
.6 to .7						115	20	214	234
.7 to .8						15	3	308	173
.8 to .9						1		257	47
.9 to 1.0								68	4
1.0 to 1.1									

A BAYESIAN APPROACH TO SECOND GUESSING "UNDECIDED" RESPONDENTS

S. James Press
University of Chicago

Chung-fang Yang
Leo Burnett Company

We are interested in the problem of estimating category probabilities in a sample survey when people are asked to answer a question with categorical answers and, for some reason, some of the respondents either answer "I don't know" or refuse to give any answer at all. Of course, for some factual questions like "What is the capital of Zambia?", a lot of people would answer "I don't know" because they really don't know. However, in some other situations, for instance, when you ask somebody his opinion about a very socially controversial question, such as, "Are you opposed to school busing?", we think that people who answer "I don't know" to this question would be very different from those people who answer "I don't know" to the previous question. We call the latter type of "I don't know's," "Undecideds." We have reasons to believe that these "Undecideds" are not really neutral.

Traditional ways of handling these "Undecided" respondents are the following three:

1) Simply to throw them out of the sample,
2) to allocate them to the unambiguous categories according to the proportions of respondents who originally, unambiguously were assigned to each of the categories,
3) to allocate them equally to each of the categories.

We think none of these methods is satisfactory theoretically. However, we do not elaborate here.

In this paper, we propose a new way of handling the problem. We assume the following model:

1) There is one question, at a time, that we are mainly interested in. We refer to it as the "main" question. We assume the main question has category responses.
2) There are some other questions which are either being asked to the respondents at the same time when the main question is asked, or which are being asked to the same group of respondents at different times. I will refer to these questions as subsidiary questions. We assume that all respondents answer the subsidiary questions unambiguously, although only some of the respondents answer the main question unambiguously (the others are the "Undecideds").
3) We assume the subsidiary questions are related to the main question, either theoretically or empirically so that they can be used to predict the respondent's answers to the main question from the way they answered these subsidiary questions.

The method of estimating the true category probabilities on the main question is a Bayes approach. It uses several types of information.

1) Subjective prior information for the category probabilities.
2) Sample frequencies for those respondents who did answer the main question unambiguously.
3) Response pattern on the subsidiary questions from the respondents who answered the main question unambiguously.

The "Undecided" respondents will be "second guessed" on the main question i.e. effectively, they will be classified into one of the unambiguous response categories on the basis of their answers to the related subsidiary questions.

Estimators of the true category probabilities on the main question can be calculated in terms of both the "decided" and second guessed "undecided" respondents. The result is the following.

Suppose n_i subjects responded unambiguously in category i of the main question, $i = 1, \ldots, M$. If m subjects are "undecided" on the question, a Bayes point estimator of the probability, q_i, that a randomly selected subject will fall into category i is given by the mean of the posterior distribution:

$$E(q_i | n_1, \ldots, n_{M-1}; z^{(1)}, \ldots, z^{(m)}) \equiv \hat{q}_i = \frac{(n_i+\alpha_i) + \sum_{j=1}^{m} P\{\Pi_i | z^{(j)}\}}{m + \sum_{j=1}^{M} (n_j+\alpha_j)}, \quad i = 1, \ldots, M, \tag{1}$$

where α_i's are parameters of the prior density for the q_i's ($\alpha_i = 1$, if we take a vague prior), $P\{\Pi_i | z^{(j)}\}$ is the marginal predictive probability for classifying the j^{th} "undecided" respondent into category Π_i of the main question, given his response on the subsidiary questions, $z^{(j)}$.

This marginal predictive probability is shown in the paper to be expressible in the form:

$$P\{\Pi_i | z^{(j)}\} = \frac{(n_i+\alpha_i)\, h(z^{(j)} | \Pi_i)}{\sum_{k=1}^{M} (n_k+\alpha_k) h(z^{(j)} | \Pi_k)} \tag{2}$$

where $h(z^{(j)} | \Pi_i)$ denotes the marginal predictive density of the response to the subsidiary questions for the j^{th} "undecided" respondent, given he belongs to category Π_i on the main question.

Therefore, $\hat{q}_i$ can be expressed as:

$$\hat{q}_i = \left[\frac{n_i + \alpha_i}{m + \sum_{j=1}^{M} (n_j + \alpha_j)} \right] \times \sum_{j=1}^{m} \left[\frac{h(z^{(j)} | \Pi_i)}{\sum_{t=1}^{M} (n_t+\alpha_t) h(z^{(j)} | \Pi_t)} \right]. \tag{3}$$

Variances for the category probability estimators

are developed in the paper, as are Bayesian credibility or confidence intervals.

The only problem remaining is to evaluate the predictive density $h(z^{(j)}|\Pi_i)$. Three cases are discussed in the paper. We discuss the cases when the $z^{(j)}$'s (i.e. the responses to the subsidiary questions) are all continuous, all discrete (categorical), and the mixed case (with some subsidiary questions having continuous responses and some having discrete responses). For illustrative purposes, we now consider explicitly the case when $z^{(j)}$ is discrete, and give the predictive density for that case.

Suppose there are q subsidiary questions with discrete type responses. Let S denote the number of possible joint responses to this set of questions. For instance, suppose there are two questions with two possible answers for each question. Then there are $S = 2 \times 2 = 4$ possible joint responses to these two questions. Let u_k, an S dimensional unit vector with a "1" in the k^{th} place and all other places are "0", denote the k^{th} joint response pattern to the subsidiary questions. Then the predictive density is given by

$$h(z^{(j)} = u_k|\Pi_i) = \frac{x_{k|i} + \delta_{k|i}}{n_i + \sum_{t=1}^{S} \delta_{t|i}}, \qquad (4)$$

where $k = 1, \ldots, S$, $i = 1, \ldots, M$, $\Delta_i = \sum_{t=1}^{S} \delta_{t|i}$, $n_i = \sum_{k=1}^{S} x_{k|i}$, $x_{k|i}$ denotes the number of respondents who answered unambiguously in category i of the main question and who were in cell k of the subsidiary questions; $\delta_{k|i}$ denotes the parameters of the natural conjugate Dirichlet prior distribution for the cell probabilities of the responses to the subsidiary set of questions; n_i is the number of respondents who anwered unambiguously in category i of the main question.

Example

Suppose that out of 100 respondents to a sensitive question, 20 people respond in each of the three possible unambiguous categories and 40 are "undecided." Moreover, suppose that there is just one subsidiary question (with four response categories) which is used, and the "decided" group on the main question respond according to the table below.

		Subsidiary Question				
		(1)	(2)	(3)	(4)	Totals
Main Question	(1)	17	1	1	1	20
	(2)	5	5	5	5	20
	(3)	1	1	1	17	20

Thus, there are 17 subjects who responded in category "1" of the subsidiary question, given they responded in category "1" on the main question, etc.

For the "undecided" group on the main question, suppose 25 respond in subsidiary category "1", and 5 respond in each of the remaining categories. Results of the analysis are given below assuming vague priors for both the q_i's and for p_i.

i	Ignoring "Undecideds"	$\hat{q}_i$	σ_{q_i}	90% credibility Interval
1	.33	.40	.04	(.34, .48)
2	.33	.34	.04	(.28, .40)
3	.33	.26	.04	(.18, .32)

Thus, if the "undecided" group had been ignored and if the q_i's were estimated on the basis of sample frequencies, column (2) would have resulted. Use of (3) yielded column (3) while column (4) gives the standard deviations. The last column was obtained by using the beta approximation described in the paper with 5% probability in each tail of the approximating distribution.

Interested readers may refer to the complete paper for further details. It is scheduled to appear in Journal of the American Statistical Association, March, 1974.

RELATIVE EFFICIENCIES OF SOME MEASURES OF ASSOCIATION FOR ORDERED TWO-WAY CONTINGENCY TABLES UNDER VARYING INTERVALNESS OF MEASUREMENT ERRORS

Charles H. Proctor, North Carolina State University

1. Introduction

As one who fairly often sees contingency tables in connection with analyzing social survey data, I recently became interested in what computations to include in the local statistical package to accompany the cross tabulation output. This paper reports on the approach that was developed to evaluate measures of association for two-way tables with ordered variates in both directions. Generally speaking, under certain conditions one measure will be best, while under other conditions another one will be. This is not an unexpected conclusion, but what may be new here is the emphasis on a somewhat more empirical method than any I have seen heretofore, in describing these conditions.

For the case of continuous variates there are results already available on the relative efficiencies of several measures of association. One major problem of such work has been to characterize, realistically and parsimoniously, the case of association or of non-independence. It has been handled by Konijn [14] and Farlie [7] in somewhat different ways. For contingency tables, the distributions suggested by Plackett [17] and referred to by Mosteller [19] would seem a natural basis for investigating power, but they were not used here. I am not aware that they have been used to exhibit relative efficiencies.

Actually, it seems to be that two approaches are being followed in the selection of a measure of association. One, employed by Kruskal [15], emphasizes that the measure computed on sample frequencies estimates a counterpart population quantity and the user should be sure he wants to know that population quantity. Interpretations are provided of these population quantities in contexts such as predicting the ordering of a pair of persons on a second IQ test from their ordering on a first IQ test. The underlying method involves discovering what actions the user wants to take when he sees his data, and then verifying that the suggested statistic will fit into that action pattern. Although the admonition to take into account the user's interests is undeniably good, the method still has its ambiguities. The amount of controversy generated, by the choice of a measure of association, among sociological methodologists is, I believe, testimony to these ambiguities [6, 16], although this issue is properly a problem in the sociology of knowledge.

Another approach is by way of measurement model theory. The investigator states what he judges to be the measurement scale of his variates, and then if, for example, both the row and column variates are of ordinal types, but not interval, he calculates the Goodman-Kruskal gamma or the Kendall tau-sub-b. At any rate, a Pearson product moment coefficient would be meaningless (a technical term [21, p. 66]) unless his scales were interval. The definitions of such distinctions among variables are most elegantly expressed as equivalence classes under certain transformations [21, p. 10]. Thus, the scale is of ordinal type if it is equivalent in distinguishing among observational units, to any other scale obtained by a monotonic transformation of it. By substituting affine transformation in place of monotonic transformation in the above, one defines an interval scale type.

The present work advocates the philosophy of this second approach, but suggests using empirical evidence in the data themselves for characterizing the measurement model type. The absence of these empirical criteria for determining scale type has always struck me as a shortcoming of the approach, and the works of Suppes and Zinnes [21, pp. 72-74] and Campbell [5, Chapters XVII and XVIII] in their final discussions of random measurement errors, serve as the stepping-off-place for the present development. The data are here viewed as sampled from a population table in which the cell probabilities are formed in part by a parent, generic stochastic process and in another part by random drift among cells arising from measurement error (also called misclassification error by Mote and Anderson [18] in this categorical variable context). The criterion to be employed for judging the measures is relative efficiency and these efficiencies are to be calculated for variations in the population table.

2. Defining the Parent Process, the Error Process and the Sampling Process

The starting point for the comparison of measures of association is a two-way contingency table, an A by B table, whose row and column variates are categorical, but ordered. The statuses of the two qualitative variables are taken to be causally symmetric, that is, jointly dependent, either on one another or on some collection of unmeasured independent variables. The underlying process of interdependence will be supposed to have produced a joint distribution of the units of observation with cell probabilities π_{ij}, with $\sum_{i=1}^{A} \sum_{j=1}^{B} \pi_{ij} = 1$. The probability π_{ij} reflects the chance that an observational unit will emerge with the $i^{\underline{th}}$ category of the row variate and the $j^{\underline{th}}$ category of the column variate. The π_{ij} will be taken to reflect the parent stochastic process.

Now further suppose that the measurement operation is to some extent fallible or that, after emergence, the unit's characteristics drift in a random fashion. This process will be called the error process. In particular, the chance of recording row category a when the unit is actually in category i will be denoted by θ_{ia}. Similarly Φ_{jb} is taken to be the chance that a unit that emerged in column category j is recorded as being in category b. Combining the two processes by supposing the errors to be independent gives the resulting probability of observing a unit in cell (a,b) as:

$$\rho_{ab} = \sum_i \sum_j \theta_{ia}\Phi_{jb}\pi_{ij} \cdot \tag{2.1}$$

If the error processes are not independent equation (2.1) becomes:

$$\rho_{ab} = \sum_i \sum_j \delta_{ij,ab}\,\pi_{ij} , \tag{2.2}$$

which may look simpler but involves many more misclassification parameters in the $\delta_{ij,ab}$'s than in the θ_{ia}'s and Φ_{jb}'s. Such a model was used by Assakul and Proctor [2] to show how great a loss of power the chi-square test suffers under measurement error. It is also presented by Hayashi [12] who corrects biases in cross tabulations by estimating the misclassification parameters.

Both notions, the one of errors and the other of the joint parent distribution, need more empirical content if they are to be usable by the data handler. Measurement errors can be examined by duplicating measurements either by using a superior or optimal measurement method or by two parallel applications of the usual technique. From the sets of duplicate determinations one can estimate the θ_{ia} and the Φ_{jb}. Discussion of this estimation problem, important as it is, would lead us too far afield. Thus we will merely suggest the model equations for measurement and drift errors that will be used to define intervalness and leave as a separate task, the estimation of the parameters in such a model.

It is taken that the misclassification probabilities follow the pattern of:

$$\theta_{ia} = (A\text{-}1)\alpha_1 e^{-\beta_1|i-a|} / \Big(\sum_{i\neq a} e^{-\beta_1|i-a|}\Big), \text{ if } i \neq a$$

$$= 1 - (A\text{-}1)\,\alpha_1 \quad \text{when } i = a \cdot \tag{2.3}$$

This may look complex but it merely states that θ_{ia} depends firstly on an overall level of error, measured by α_1. If β_1 is large this error occurs largely in adjacent categories, while if β_1 is small then distant categories may be confused. Thus, if α_1 is zero, there is no measurement nor drift error, while if β_1 is zero the pattern is of "ordinal" type, and the larger β_1 becomes the more the pattern can be called "interval".

To complete the picture the Φ_{jb}'s may be similarly defined by:

$$\Phi_{jb} = (B\text{-}1)\alpha_2 e^{-\beta_2|j-b|} / \Big(\sum_{j=b} e^{-\beta_2|j-b|}\Big), \text{ if } j \neq b$$

$$= 1 - (B\text{-}1)\,\alpha_2 \quad \text{when } j = b \cdot \tag{2.4}$$

However, in all of the following calculations we have taken $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$.

Having introduced the generic or parent process represented by π_{ij} and the drift or measurement error process by the θ_{ia}'s and Φ_{jb}'s, it remains to specify the sample selection process. If one supposes that each of the n observational units had the same chance, namely π_{ij}, of being in the (i,j) cell, and was subject to the same error process and that these processes acted independently from one unit to the next, then the cell frequencies, the n_{ab}, would be distributed in accord with a multinomial distribution with AB classes having underlying probabilities $\{\rho_{ab}\}$. And this is what we will assume. Of the three sampling distributions cases distinguished by Barnard [4] as (a) both margins fixed, (b) one fixed, and (c) both free, ours is the third.

3. Measures of Association to be Compared

The four principal measures of association to be examined are the Goodman-Kruskal G, Kendall's TB, the Kendall-Stuart TC, and a coefficient R that is a Pearson product-moment correlation coefficient using integer row and column scores. The first three appear to be widely used in the social sciences, while the fourth is congenial to the author's naive numerical point of view and follows Williams' [22] and Yates' [23] approaches. The definitional formulas are (see [11, p. 325; 10, p. 751; and 13, p. 563]):

$$G = (P_s - P_d)/(P_s + P_d) \tag{3.1}$$

$$(3.2)\quad TC = (P_s - P_d)[A/(A-1)]$$

$$(3.3)\quad TB = (P_s - P_d)/[(1 - \Sigma R_{a.}^2)(1 - \Sigma R_{.b}^2)]^{1/2}$$

$$(3.4)\quad R = \sum_{a=1}^{A}\sum_{b=1}^{B}(a - \bar{a})(b - \bar{b})\, R_{ab}/s_a s_b ,$$

where the quantities involved in these formulas are defined as:

$$P_s = 2\sum_a\sum_b R_{ab}\{\sum_{a'>a}\sum_{b'>b} R_{a'b'}\} ,$$

$$P_d = 2\sum_a\sum_b R_{ab}\{\sum_{a'>a}\sum_{b'<b} R_{a'b'}\} ,$$

$$R_{a.} = \sum_{b=1}^{B} R_{ab} \quad \text{and} \quad R_{.b} = \sum_{a=1}^{A} R_{ab} ,$$

with

$$s_a^2 = \sum_{a=1}^{A}(a - \bar{a})^2 R_{a.}$$

and

$$s_b^2 = \sum_{b=1}^{B}(b - \bar{b})^2 R_{.b} ,$$

where

$$\bar{a} = \sum_a a\, R_{a.} \quad \text{and} \quad \bar{b} = \sum_b b\, R_{.b} .$$

4. Basis of Comparison, Relative Efficiency

The choice of criteria for comparing one measure with the other is not so much a mathematical question nor an empirical one, it is more like a moral one. In accord with cannons of argument from tradition in this ethical field, I will adduce that the criterion, namely efficiency, that I will use is widely accepted in the statistical profession. In comparing two procedures one frequently computes the ratio of the two sample sizes required to attain the same variance in estimating some parameter or required to achieve the same power in testing some hypothesis. This ratio is called relative efficiency and since larger samples are generally more costly, it shows which procedure will, in that sense, make best use of the observations. In the present case it is somewhat difficult to decide whether the problem is one of estimating a parameter or of testing an hypothesis.

As defined, the four measures are statistics; they are functions of the sample relative frequencies, the R_{ab}'s. In each case there is the same function of the population relative frequencies, the ρ_{ab}'s, that constitutes a population parameter. However, there are four different parameters, and it would seem necessary first to decide which parameter was needed (as Kruskal [16] does) and then use the corresponding statistic.

The alternative point of view we have adopted is that the null hypothesis, H_0 is, in all cases, that of the independence of row and column categories, while the investigator is interested in detecting departures from H_0. From his knowledge, experience and perhaps a glance at the data, he suggests a structure on the π_{ij} and supposes some pattern of θ_{ia} and Φ_{jb}. We thus arrive at a set of ρ_{ab}, the alternative hypothesized population values. The sample is assumed drawn from this population and the measure is computed. It is proposed to compare any two such measures by the ratio of sample sizes that would be required to attain the same power at the alternative hypothesized values.

The reasoning goes that each statistic, being a function of cell relative frequencies, is asymptotically, as sample size increases, normally distributed (about zero if H_0 holds) and one would divide the statistic by its standard deviation to obtain a critical ratio. He would then refer this Z-value, say, to a table of areas under the normal curve to furnish a significance probability for rejecting H_0. If the test is made at a 5% level of significance then a distance of Z = 1.960 (where $\Phi(1.960) = .975$ and Φ is the normal distribution function) away from zero would lead to rejecting H_0. In order to assure this rejection with fairly high probability of, say, .80 would require the population mean Z-value to be 1.960 plus .842 (since $\Phi(.842) = .80$), or 2.802 away from zero. The population mean Z-value is the parameter value under the alternative hypothesis divided by the standard deviation. In all cases, the standard deviations of the statistics, based as they (the statistics) are upon sample proportions, are, to a first approximation, proportional to $n^{-1/2}$. Thus the required sample sizes become (note that $2.802^2 = 7.85$):

$$(4.1)\quad n_G = 7.85\, V_G/\gamma^2, \quad n_{TC} = 7.85\, V_{TC}/\tau_c^2 ,$$

$$n_{TB} = 7.85\, V_{TB}/\tau_b^2, \quad n_R = 7.85\, V_R/\rho^2$$

where the variances are given as $V(G) = V_G/n$, $V(TC) = V_{TC}/n$, etc. and γ^2, τ_c^2, τ_b^2 and ρ^2 are given by (4.1) to (4.4) with ρ_{ab} in place of R_{ab}. When comparing these sample sizes, one sees that ratios such as γ^2/V_G, τ_c^2/V_{TC}, etc. are the crucial quantities. These ratios will be denoted as the "precisions" of the statistics with notation:

$$(4.2)\quad P_G = \gamma^2/V_G, \quad P_c = \tau_c^2/V_{TC},$$

$$P_B = \tau_b^2/V_{TB}, \quad \text{and} \quad P_R = \rho^2/V_R .$$

The argument here is closely akin to that of Bahadur [3] in which his quantity c, or "slope", is equal to γ^2/V_G or τ_c^2/V_{TC}, and so forth.

5. Variance Expressions and Derivations

In the case of G, Goodman and Kruskal [11] give

$$(5.1)\quad V_G = nV(G) = 16[\Pi_s^2\Pi_{dd} - 2\Pi_s\Pi_{sd} + \Pi_d^2\Pi_{ss}]/(\Pi_s + \Pi_d)^4 ,$$

where

$$\Pi_{ss} = \sum_a \sum_b \rho_{ab}\Big[\sum_{a'>a}\sum_{b'>b}\rho_{a'b'} + \sum_{a'<a}\sum_{b'<b}\rho_{a'b'}\Big]^2 ,$$

$$\Pi_{dd} = \sum_a \sum_b \rho_{ab}\Big[\sum_{a'>a}\sum_{b'<b}\rho_{a'b'} + \sum_{a'<a}\sum_{b'>b}\rho_{a'b'}\Big]^2 ,$$

and

$$\Pi_{sd} = \sum_a \sum_b \rho_{ab}\Big[\sum_{a'>a}\sum_{b'>b}\rho_{a'b'} + \sum_{a'<a}\sum_{b'<b}\rho_{a'b'}\Big] \times \Big[\sum_{a'>a}\sum_{b'<b}\rho_{a'b'} + \sum_{a'<a}\sum_{b'>b}\rho_{a'b'}\Big] ,$$

and where Π_s and Π_d are the counterparts of P_s and P_d in formulas (4.5) and (4.6) when ρ_{ab} is used in place of R_{ab}. Formula (5.1) can be derived by a straight forward, if tedious, application of a method described by R. A. Fisher [8, p. 309-310] for the variance of any statistic, T say, when T is a function of frequencies that obey a multinomial distribution. Fisher's formula is:

$$(5.2)\quad \frac{1}{n}V(T) = \sum_a \sum_b \{N_{ab}(\frac{\partial T}{\partial N_{ab}})^2\} - [\frac{\partial T}{\partial n}]^2\Big|_{N_{ab}=n\rho_{ab}} .$$

The last term is zero for three of the four statistic since n enters explicitly only into TB. For the other three, the formula can be written in terms of relative frequencies as:

$$(5.3)\quad nV(T) = \sum_a \sum_b R_{ab}(\frac{\partial T}{\partial R_{ab}})^2 .$$

The results are:

$$(5.4)\quad V_{TC} = (\frac{A}{A-1})^2 4[\Pi_{ss} - 2\Pi_{sd} + \Pi_{dd} - (\Pi_s - \Pi_d)^2] .$$

After quite messy algebra it turns out that:

$$(5.5)\quad V_{TB} = \frac{4}{\Delta_a\Delta_b}\{(\Pi_{ss} - 2\Pi_{sd} + \Pi_{dd}) + (\Pi_s - \Pi_d) \times \sum_a \sum_b \rho_{ab}(\sum_{(ab)}^{(s)}\rho_{a'b'})(\frac{\rho_{.b}}{\Delta_b} + \frac{\rho_{a.}}{\Delta_a}) + \frac{1}{4}(\Pi_s - \Pi_d)^2 \sum_a \sum_b \rho_{ab}(\frac{\rho_{.b}}{\Delta_b} + \frac{\rho_{a.}}{\Delta_a})^2\} - \frac{(\Pi_s - \Pi_d)^2}{\Delta_a\Delta_b}[\frac{1}{\Delta_b} + \frac{1}{\Delta_a}]^2 ,$$

where $\Delta_a = 1 - \sum \rho_{a.}^2$ and $\Delta_b = 1 - \sum \rho_{.b}^2$, while the result for R is:

$$(5.6)\quad V_R = \{(1 + \rho^2/2)\sum_a \sum_b (a-\bar{a})^2(b-\bar{b})^2\rho_{ab} - \rho\sigma_a \sum_a \sum_b (a-\bar{a})(b-\bar{b})^3\rho_{ab}/\sigma_b - \rho\sigma_b \sum_a \sum_b (a-\bar{a})^3(b-\bar{b})\rho_{ab}/\sigma_a + \frac{\rho^2}{4}(\sigma_a^2\sum_b (b-\bar{b})^4\rho_{.b}/\sigma_b^2 + \sigma_b^2\sum_a (a-\bar{a})^4\rho_{a.}/\sigma_a^2)\}/\sigma_a^2\sigma_b^2 .$$

A comment may be in order on the status, as approximations, of the quantities V_G, V_{TC}, V_{TB} and V_R. The random variables G, TC, TB and R have distributions induced, as mentioned before, by the multinomial distribution of the R_{ab}. Their variances are complicated functions of the ρ_{ab}; in fact, except for TC, the expressions are infinite series, the terms of which can be collected in increasing powers of n^{-1}. The expressions for V_G, V_{TC}, V_{TB}, and V_R are the coefficients of n^{-1} in these series. As n increases the other terms become smaller at a rate faster than this first term. Basing our comparison on this first term implies that our results hold only for large sample sizes.[2]

6. Relative Efficiency of the Chi-Square Contingency Table Test

In order to complete the picture, the chi-square test statistic for contingency tables has been included in the comparisons of efficiencies. The statistic is:

$$(6.1)\quad X^2 = n\sum_a \sum_b (R_{ab} - R_{a.}R_{.b})^2/R_{a.}R_{.b} .$$

The, so called, non-centrality parameter is:

$$(6.2)\quad \lambda = \sum\sum(\rho_{ab} - \rho_{a.}\rho_{.b})^2/\rho_{a.}\rho_{b.}$$

By an argument similar to that which led to formulas (4.1) only based on the non-central chi-square distribution, the formula for required sample size is:

$$(6.3)\quad n_{X^2} = \lambda(DF, .05, .80)/\lambda .$$

Here $\lambda(DF, .05, .80)$ is an entry from tables by E. Fix [9], which is the value of the non-centrality parameter for the non-null distribution of X^2 such that a test at $\alpha = .05$ based on X^2 will reject H_0 with probability .80. In particular, $\lambda(25, .05, .80) = 22.843$ is the value we will use when discussing some 6 by 6

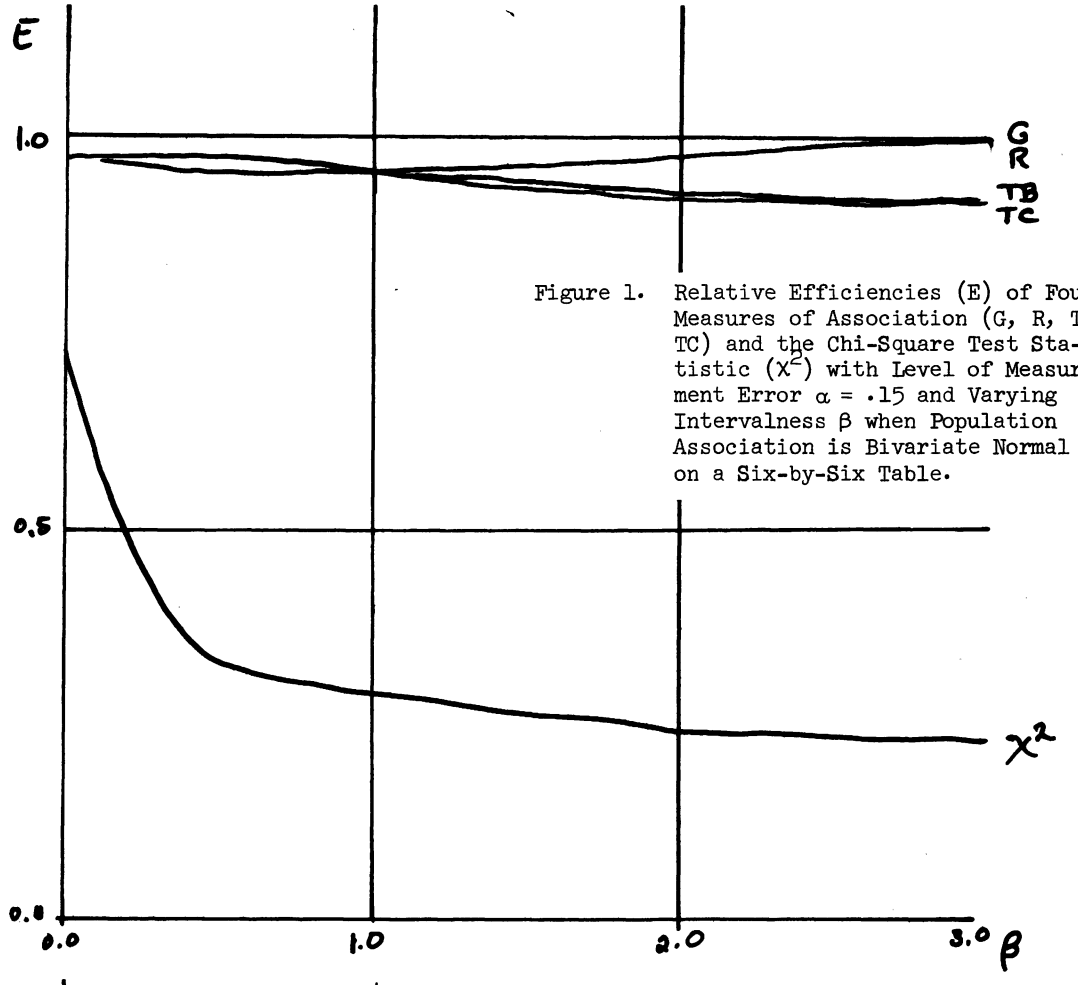


Figure 1. Relative Efficiencies (E) of Four Measures of Association (G, R, TB, TC) and the Chi-Square Test Statistic (X^2) with Level of Measurement Error $\alpha = .15$ and Varying Intervalness β when Population Association is Bivariate Normal on a Six-by-Six Table.

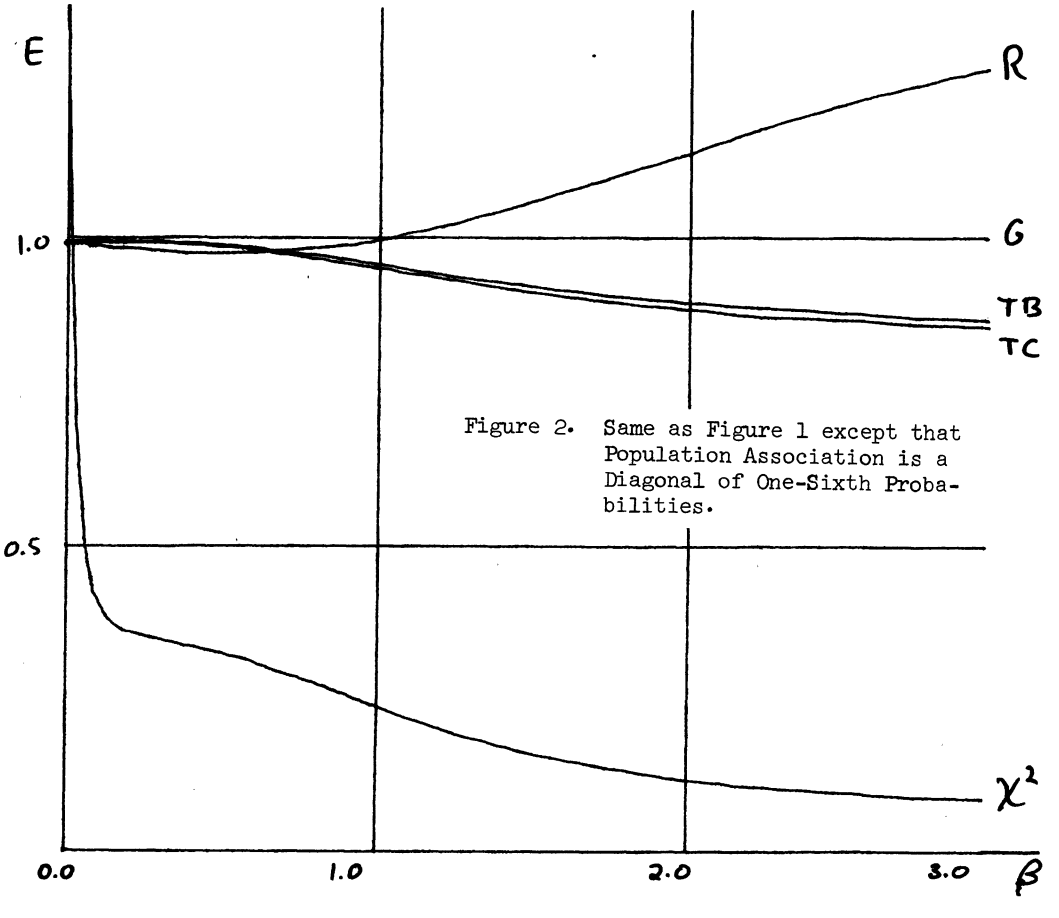


Figure 2. Same as Figure 1 except that Population Association is a Diagonal of One-Sixth Probabilities.

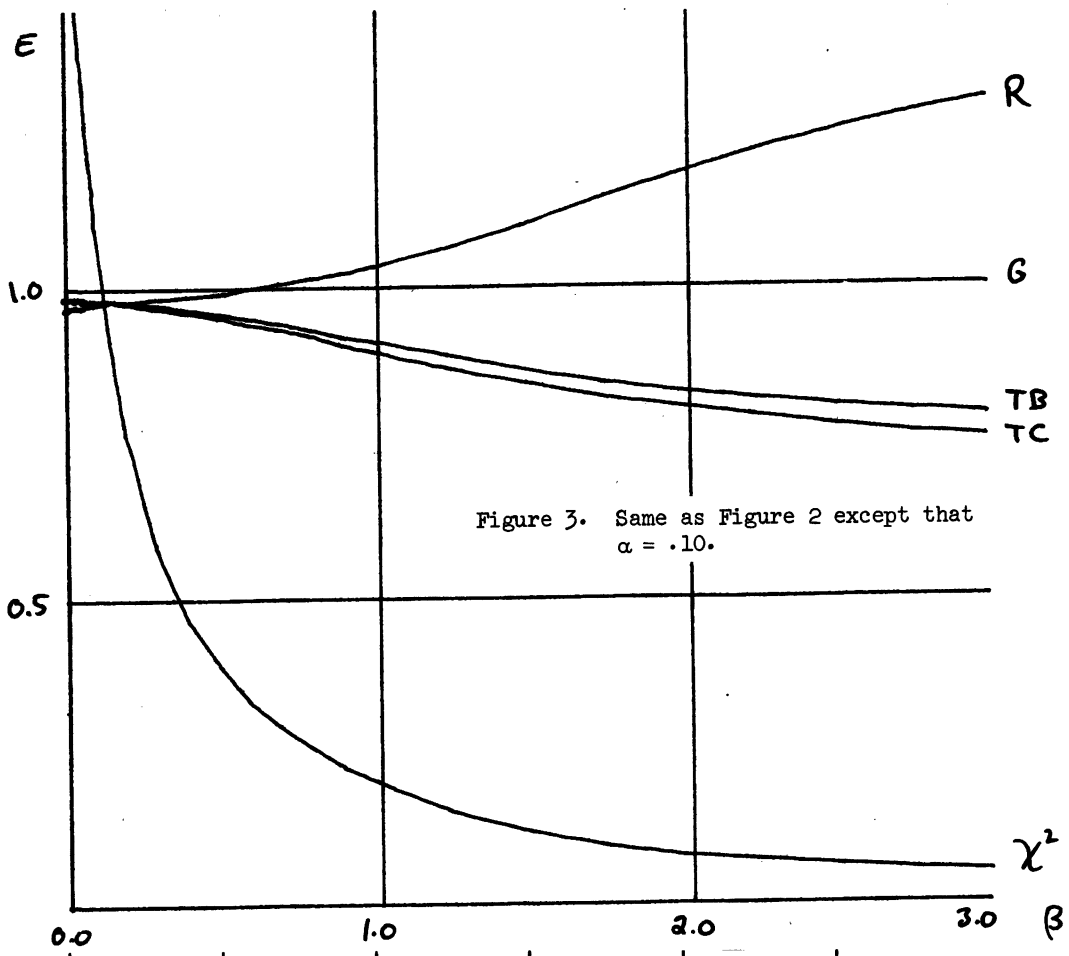


Figure 3. Same as Figure 2 except that $\alpha = .10$.

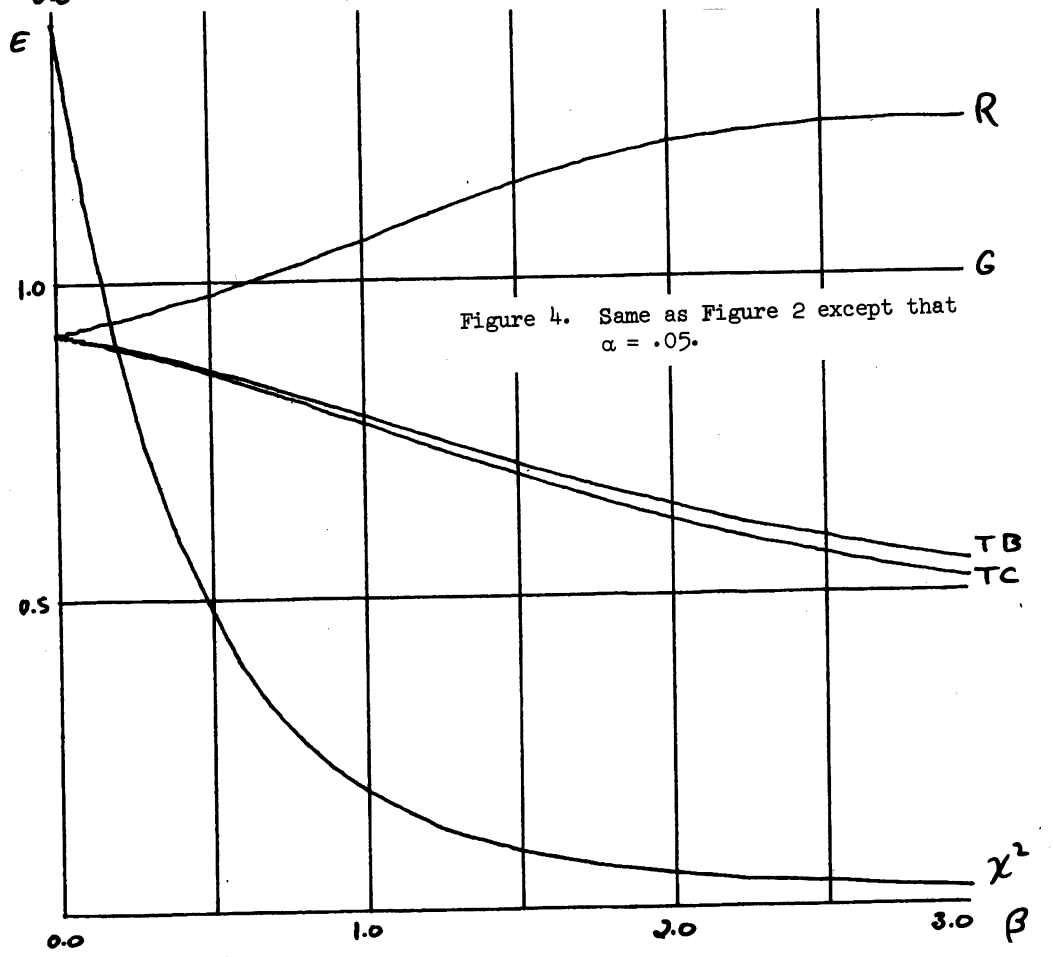


Figure 4. Same as Figure 2 except that $\alpha = .05$.

tables below. As a precision for the chi-square test to be put into (5.5) we will use:

$$(6.4)\quad P_{X^2} = 7.85\ \lambda/\lambda(DF, .05, .80)\ .$$

7. Numerical Results on Efficiencies

In presenting results on relative efficiencies the coefficient G has been taken as a base of comparison and its efficiency set at 1.000. The relative efficiencies of the other statistics become $E_{TC} = P_C/P_G$, $E_{TB} = P_B/P_G$, $E_R = P_R/P_G$ and $E_{X^2} = P_{X^2}/P_G$. For the values of the π_{ij}, probabilities under the joint normal distribution were used in the first results, and A and B were both taken as six. The π_{ij} were probabilities in a joint distribution with correlation coefficient equal to .80 and the marginal probabilities were all set equal to one-sixth.[3] The matrix of the π_{ij} thus appears as:

$$(\pi_{ij}) = \begin{pmatrix} .106 & .031 & .011 & .010 & .004 & .004 \\ .031 & .074 & .034 & .018 & .005 & .004 \\ .011 & .034 & .063 & .031 & .018 & .010 \\ .010 & .018 & .031 & .063 & .034 & .011 \\ .004 & .005 & .018 & .034 & .074 & .031 \\ .004 & .004 & .010 & .011 & .031 & .106 \end{pmatrix}.$$

If there is no error, that is $\rho_{ab} = \pi_{ij}$, then the following are the efficiencies:

$$E_G = 1.000,\ E_{TC} = .825,\ E_{TB} = .819$$

$$E_R = 1.006,\ E_{X^2} = .271\ .$$

One concludes for this case of exact measurement of an interval-type association that R, G, TB and TC have nearly the same sensitivity to such an alternative hypothesis, while chi-square is much less sensitive.

As measurement or drift error is introduced by way of the parameter α, we find, in general, that R, TB and TC drop in efficiency relative to G when we maintain the underlying joint normal distribution as a source of π_{ij} values. Figure 1 shows this for $\alpha = .15$. Notice how similar are TC and TB, as might be expected. Also notice how R recovers its efficiency as intervalness of the error pattern increases, i.e., as β increases.

When the error-free distribution is taken to be a diagonal of one-sixth probabilities then the results shown in Figure 2 appear. In this case the coefficient R becomes the most efficient as intervalness increases. TB is a bit more efficient than TC but both are somewhat below G. The pattern of results in Figure 3 can be accentuated by reducing the level of error from $\alpha = .15$ to $\alpha = .10$ and $\alpha = .05$. These results are shown in Figures 4 and 5.

8. Conclusions

A major conclusion is that, while variations in the error process and in the underlying pattern of association will lead the relative efficiencies to change a bit, the four measures of association are all quite similar. This seems to suggest that in the absence of any clear-cut superior measure of association the data analyst should use whichever measure he favors. However, he should direct some attention to the measurement error process. There should be re-testing and re-coding done so as to learn between which categories is there misclassification. Then this source of uncertainty can be adjusted for in estimating the pattern of association among the true-category variates. Also, when working with qualitative data, one should consider formulating a realistic probability model for the cell frequencies and then estimating the parameters of that model rather than estimating some conventionally employed measure of association.

If, however, a measure of association must be recommended then my judgement would be the following. In so far as it seems reasonable that a numerical variate will be defined some day and the present categories will become ranges of this variate, then use scores and compute R. If not, then use G when two variates are involved, particularily if it seems that the measurement errors follow a relatively "flat" pattern with a small β. Kendall and Stuart [13, p. 566] report results suggesting the superiority of G over TC and conclude that: "If this is shown to be true in general, this fact ... would make [G] likely to become the standard measure of association for the ordered case." Our calculations show it to be true fairly generally. If more than two categorical variates with measurement error pattern having small β are being analyzed then one might use TB. This is because TB is a bona fide correlation coefficient, albeit of pair data. Consequently, its correlation matrix will be positive semi-definite and so lends itself to multivariate calculations.

FOOTNOTES

1. Although G, TB, and R are ratios of random variables, TC is not; and it is possible to obtain an expression for the exact variance of TC as:

$$(5.4a)\quad V(TC) = \left(\frac{A}{A-1}\right)^2 \frac{4}{n-1} \times [\Pi_{dd} - 2\Pi_{sd} + \Pi_{dd} - (\Pi_s - \Pi_d)^2] + \frac{2}{n(n-1)} [\Pi_s + \Pi_d - 4(\Pi_{dd} - 2\Pi_{sd} + \Pi_{ss}) + 3(\Pi_s - \Pi_d)^2]$$

When n is of moderate size, one would multiply TC by (n-1)/n to obtain the quantity t_c in

[13, p. 563], and then the variance expression (5.4a) would be multiplied by $[(n-1)/n]^2$.

2. It may be of interest to note how the proportional importance of the extra part of formula (5.4) to the part in 1/n decreases. For a case with $\tau_c = .247$ in a 6 by 6 table the exact variance was found to be $.42878/n-1 + .88073/n(n-1)$. This suggests that for moderate sample sizes our approximation would not be too good, but for large sample sizes the difference is negligible. With n = 20, the increase of exact over approximate variance is 9.3% and for n = 100 it drops to 2.0%.

3. These calculations were made by Anne Dalhouse using the charts provided in the National Bureau of Standard's Handbook [1, pp. 936].

REFERENCES

[1] Abramowitz, M. and Stegun, I. A. (editors). Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematics Series, No. 55, Washington, D. C.: U. S. Government Printing Office, 1964.

[2] Assakul, K. and Proctor, C. H. "Testing Independence in Two-way Contingency Tables with Data Subject to Misclassification," Psychometrika, 32 (March, 1967), 67-76.

[3] Bahadur, R. R. "Stochastic Comparison of Tests," The Annals of Mathematical Statistics, 31 (June, 1960), 276-295.

[4] Barnard, G. A. "Significance Tests for 2 × 2 Tables," Biometrika, 34 (January, 1947), 123-

[5] Campbell, N. R. Foundations of Science: The Philosophy of Theory and Experiment, New York: Dover Publications, Inc., 1957.

[6] Costner, H. L. "Criteria for Measures of Association," American Sociological Review, 30 (June, 1965), 341-353.

[7] Farlie, D. J. G. "The Asymptotic Efficiency of Daniels' Generalized Correlation Coefficients," Journal of the Royal Statistical Society, Series B 23, No. 1 (1961), 128-142.

[8] Fisher, R. A. Statistical Methods for Research Workers, 13th edition-revised, New York: Hafner Publishing Company, Inc., 1958.

[9] Fix, E. "Tables of Noncentral χ^2," University of California Publication in Statistics, 1 (1949), 15-19.

[10] Goodman, L. A. and Kruskal, W. H. "Measures of Association for Cross Classification, II: Further Discussion and References," Journal of the American Statistical Association, 54 (March, 1959), 123-163.

[11] Goodman, L. A. and Kruskal, W. H. "Measures of Association for Cross Classification III: Approximate Sampling Theory," Journal of the American Statistical Association, 58 (June, 1963), 310-364.

[12] Hayashi, C. "Response Errors and Biased Information," Annals of the Institute of Statistical Mathematics," 20 (No. 2, 1968), 24-28.

[13] Kendall, M. G. and Stuart, Alan. The Advanced Theory of Statistics: Vol. 2 Inference and Relationship, New York: Hafner Publishing Company, 1961.

[14] Konijn, H. S. "On the Power of Certain Tests for Independence in Bivariate Populations," The Annals of Mathematical Statistics, 27 (June, 1956), 300-323.

[15] Kruskal, W. H. "Ordinal Measures of Association," Journal of the American Statistical Association, 53 (December, 1958), 814-861.

[16] Morris, R. N. "Multiple Correlation and Ordinally Sealed Data," Social Forces, (March, 1970), 299-311.

[17] Plackett, R. L. "A Class of Bivariate Distributions," Journal of the American Statistical Association, 60 (June, 1965), 516-522.

[18] Mote, V. L. and Anderson, R. L. "The Effect of Misclassification on the Properties of χ^2 - Tests," Biometrika, 52 (June, 1965), 95-109.

[19] Mosteller, F. "Association and Estimation in Contingency Tables," Journal of the American Statistical Association, 63 (March, 1968), 1-28.

[20] Rao, C. R. Linear Statistical Inference and Its Applications, New York: John Wiley and Sons, Inc., 1965.

[21] Suppes, P. and Zinnes, J. L. "Basic Measurement Theory," Chapter 1 in Luce, R. D., Bush, R. R. and Galanter, E. (editors), Handbook of Mathematical Psychology, Vol. 1, New York: John Wiley and Sons, Inc., 1963.

[22] Williams, E. J. "Use of Scores for the Analysis of Association in Contingency Tables," Biometrika, 39 (December, 1952), 274-289.

[23] Yates, F. "The Analysis of Contingency Tables with Groupings Based on Qualitative Characters," Biometrika, 35 (May, 1948), 176-181.

NATIONAL ESTIMATES OF THE PREVALENCE OF NARCOTIC ADDICTION

Alex Richman, with the assistance of Vincent V. Richman
Department of Psychiatry
Beth Israel Medical Center, and Mount Sinai School of Medicine of the
City University of New York

INTRODUCTION

Knowledge of trends in the extent of narcotic addiction is essential for the evaluation and planning of treatment and prevention programs. Various attempts to estimate prevalence and incidence have been characterized by differences in definition, nomenclature, assumptions and methodology. (Richman, 1974).

This paper describes some of the statistical fallacies associated with recent national estimates of the prevalence of narcotic addiction. The log normal nature of the spatial distribution of narcotic addiction is identified and a new biometric approach for estimating prevalence is illustrated by recent data for New York City.

BACKGROUND

Population surveys of the extent of narcotic addiction have been unsatisfactory because of difficulties in surveying those at highest risk of narcotic use. Chambers concluded that his 1971 interview survey of New York City residents was successful in identifying only "one-third" of the actual population of regular heroin users.

The National Commission on Marijuana and Drug Abuse found 0.2% of youths and 0.1% of adults had used heroin within six months of interview but emphasized the limitations of such data in providing no indication of the extent or characteristics of users who go beyond the initial experience to adopt patterns of more prolonged, frequent or intensive drug use, for which data on duration, frequency, and intensity are essential.

SOURCES OF DATA

Information about "known" users has been derived from law enforcement agencies, clinical treatment facilities, and cumulated from various sources by case registers.

Blumstein et al emphasized that these data sources are generally viewed as incomplete in coverage, as unpurged of the dead, the cured, and the emigrant, and as the products of recording processes listing individuals, at times, without due regard for evidence of drug use or addiction; they concluded that attempts to estimate the national extent of heroin addiction were "...relying on questionable extrapolation, based on tenuous hypotheses, from elusive data..."

Law Enforcement Agencies

Since 1954 the Bureau of Narcotics and Dangerous Drugs has published annual reports of the number of "active" narcotic users in the United States. The report for 1969 acknowledged that BNDD was unable to assess the degree of validity of their statistics for determining prevalence of narcotic use.

Clinical Treatment

A second source of data on heroin users is derived from the experience of specific clinical treatment facilities. These clinical data are usually affected by the treatment modalities and admission policies of the specific agencies and do not represent the incidence or prevalence of heroin dependency in the community.

Case Registers

The New York City Narcotics Register collates reports required by law from health agencies, law enforcement groups, private physicians, and other persons, institutions and agencies, having contact with drug abusers. No standard definition of narcotic dependency is used. All reports are accepted without confirmation.

Narcotic Register tabulations of the cumulated number reported over a period of time are often misinterpreted as showing the characteristics of persons alive, residing in the community and addicted at the end of that time period. The Narcotics Register (1973) now recommends that at least 25% of all first reports should not be considered in estimating prevalence and in addition currently assumes an annual inactivation rate of somewhere between 10% and 17%.

Error is further compounded when Narcotics Register data are used without correction for the effect of aging. Eighty-eight thousand drug abusers were reported under the age of 25 to the Narcotics Register by the end of 1970; if all of these were still addicted in New York City at the end of 1970, it is estimated that thirty per cent would be over 25 at the end of 1970. This lack of aging (the Peter Pan Principle) frequently results in comparison groups being erroneously considered to be older than those reported to the Narcotics Register.

- - - - - -

Reprint requests should be addressed to 307 Second Avenue, New York 10003

NATIONAL ESTIMATES OF PREVALENCE AND INCIDENCE

The extent of narcotic addiction has been estimated in many ways. This diversity of methods results from the lack of any single, satisfactory approach. In addition to inconsistent definitions of prevalence, a wide variety of ratios are used, without justification to "correct" for undercounting. (U.S. Senate Committee), (Richman, 1974)

Combinations of Data from BNDD and New York City Narcotics Register

McGlothlin et al estimated there were 375,000 narcotic addicts in the United States at the end of 1971. They assumed that 125,000 of the 175,000 names of the New York City Narcotics Register were active narcotic addicts, and that the Register was 60-80% complete--resulting in a prevalence of 150-200 thousand. Since 40% of the 82,000 persons listed on the 1971 BNDD file were from New York City, the estimate for New York City was inflated to a national estimate of 375,000.

Drug Related Deaths

The number of "heroin overdose" deaths has been used as a multiplier to estimate prevalence. Dupont estimated that 42 deaths from opioid overdosage in Washington, D.C. represented 16,800 heroin users. Such extrapolations are unjustified not only because of the problem in selecting an appropriate multiplier and the difficulty in defining drug related deaths (Brecher), but because there are major changes in mortality among drug users followed over a number of years. (Jackson and Richman)

Andima et al determined that 61% of the 114 deaths aged 15-19 in 1970 were previously unknown to the Narcotics Register and a multiplier of 61/39 was applied to the 36,019 Narcotics Register cases first reported before the age of 20 up to the end of 1970.

This method assumes that all persons reported under the age of 20 to the Narcotics Register before 1971 were still alive and addicted in New York City and under 20 years of age at the end of 1970. Applied to all age groups, this approach produced an estimate of 316,918 New York City narcotic addicts at the end of 1970.

Crime Statistics

Newmeyer has estimated prevalence by making multiple assumptions of the number of unreported burglaries, the proportion of burglaries committed by opiate addicts the cost of the average addict burglar's habit while using opiates, and the proportion of addicts whose habits are entirely supported by burglary,etc.

Capture-Recapture Techniques

Greenwood used capture-recapture methods for estimating the prevalence of heroin addiction. BNDD data were used to derive 1969 estimates of 150,000 heroin addicts for New York City, and a national total of 315,000. Later calculations provided national estimates of 524,000 for 1970 and 559,000 for 1971.

REQUIREMENTS FOR PREVALENCE ESTIMATES

All of the above methods involve assumptions which are not supportable. Sources are used which do not reflect current clinical activity - Narcotics Register counts include persons first reported before 1964 and BNDD lists include persons reported up to five years earlier. Addiction is assumed by some methods to be uniformly permanent, without remission, to be associated with permanent residence in New York City, and to involve a suspension of aging.

The natural history of narcotic addiction is not uniform. At the end of 1967 there were more active addicts (22,535 addicts) known to BNDD for over five years than there were three years later at the end of 1970 (20,596). (Statistical Abstract of the United States) Robins has demonstrated the low degree to which narcotic addiction persisted among veterans found to be using narcotics in Vietnam. The clinical characteristics and treatment needs of addicts from different report sources are erroneously assumed to be similar.

Finally, demographic concentrations are not recognized in that national extrapolations are often made from New York City data without consideration of the marked demographic differentials.

The rest of this paper describes an approach which meets the above requirements. This approach includes:

1) identification of the log-normal nature of the spatial distribution of narcotic addiction in New York City.
2) estimation of geometric dispersion of the prevalence of narcotic addiction derived from Narcotics Register data.
3) use of data from a clinical facility with enhanced opportunity for contact with actively addicted persons in a defined geographic area and calculation of age-sex-color specific population-based ratios.

4) application of order statistics for estimating the city-wide prevalence of addicts.

SPATIAL DISTRIBUTION OF NARCOTIC ADDICTION

Narcotic addiction is highly concentrated within certain geographic areas. Spatial concentrations of drug addiction were shown in Chicago by Faris and Dunham, and Dai.

Chein et al studied the distribution of 3,457 boys aged 16-20 reported from New York City law enforcement agencies and hospitals as involved with narcotics, between 1949 and 1955. Over three-fourths of the census tracts had no cases reported. Census tracts with 29% of the boys contributed 83% of the cases. Koval's analysis of New York City Narcotics Register data found similar spatial concentrations for 1964-1967.

STATISTICAL DISTRIBUTION OF PREVALENCE RATIOS

This section demonstrates the log normal distribution of the prevalence of narcotic addiction within the 30 Health Districts of New York City. Koval derived the age-specific ratios of narcotic addicts within the 30 New York City Health Districts from reports made to the Narcotics Register 1964-1967. Koval's data were plotted on log normal probability paper for the group aged 25-29 and a straight line fitted by inspection by procedures described by King and Ferrell. See Figure .

The intersection of the fitted straight line with the 50% probability line is a good estimate of the mean of the logarithms or the geometric mean- in this case 2.1% for the population aged 25-29. The standard deviation of the logarithms or geometric dispersion was estimated by dividing the value at the 93.3% point (100) by the value at the 6.7% point (4) and taking the cube root of the result. The estimated geometric dispersion or g is 2.9. Now g is used similarly to σ for the normal distribution with the substitution of multiplication and division for addition and subtraction. The geometric mean multiplied by $g \pm 2$ includes about 95% of the values. From Koval's Narcotics Register data, 95% of the values are included within the range 0.2%-17.7%, that is the 95% confidence limits for the prevalence of narcotic addiction are estimated to be between 0.2% and 17.7% for the group aged 25-29 years.

Confidence intervals for this distribution were determined, the geometric dispersion being raised to a power appropriate for both percentile and sample size of 30 (King). All of the plotted points fell within the confidence intervals. The 95% confidence limits of the estimated geometric mean of this distribution are 1.4% to 3.1%.

The 1970 Census enumerated 596,566 persons in the 25-29 age group. Applying the estimated geometric mean of 1.4% - 3.1% to this population gives a 95% confidence interval of 8,400-18,600 addicts aged 25-29.

A second source of data on the recent spatial distribution of narcotic addiction was found in the census tract distribution of 833 persons classified as narcotic addicts by the Baltimore City Police Narcotic Unit from December 1, 1966 to November 30, 1968. (Nurco) The median rate was 77 per 100,000 and ranged from one per cent to zero. The transformation of log (x + 20) was plotted, a straight line fitted, and estimates made for the geometric mean of 90 per 100,000 and a geometric dispersion of 2.2.

PROBLEMS IN USE OF NARCOTICS REGISTER DATA

We have referred earlier to problems in defining current clinical characteristics and needs of persons reported to the Narcotics Register from various sources over a period of time.

There is an additional problem in selecting denominators appropriate for calculating geographic distribution of heroin dependency. With longer time intervals, there is a progressively increasing discrepancy between the number of persons who have lived in the area at any time during that interval and the number at the midpoint of that interval (persons and person-years).

In areas with many addicts, the cumulated number reported over many years can be a relatively high proportion of the census population. Nearly "15%" of Cental Harlem's population aged 15-44 were reported to the Narcotics Register between 1964-1967 (Koval).

The Narcotics Register data do not adequately describe the geographic distribution of persons currently defined as actively addicted. It is necessary to have data on the characteristics and geographic distribution of persons known to be dependent on narcotics, which give a more current picture of actively addicted persons during a relatively short time period. Such data are described in the next section.

M.J. BERNSTEIN INSTITUTE, BETH ISRAEL MEDICAL CENTER, NEW YORK

These data are from narcotic addicts seen at M.J. Bernstein Institute, MJBI (formerly the Manhattan General Hospital) of the Beth Israel Medical Center, whose need for care has been medically substantiated. (Richman, Feinstein and Trigg)

Although MJBI is located in Lower Manhattan, patients come from all Boroughs of New York City. The relation of MJBI patients to addicts seen elsewhere has been assessed for a sample of 155 persons first reported to the Narcotics Register in 1967. By the end of 1968, 25% of the Narcotics Register sample had been admitted to MJBI. (Richman et al, 1971)

Andima reports that one-half of all addicts who died in New York City in 1970 had been known to the Narcotics Register before death. One-half of those reported to the Narcotics Register before death, had been previously hospitalized at MJBI. (Jackson and Richman). Thus, MJBI had previous contact with 25% of narcotic-related deaths during 1970.

POPULATION BASED PREVALENCE RATIOS, LOWER EAST SIDE HEALTH DISTRICT, MANHATTAN

Narcotic addicts who applied to MJBI for care or were in treatment during June 1, 1970 - June 30, 1972 completed a questionnaire equivalent to that used in the 1970 Census. (Richman, 1973) About three thousand individual residents of the Lower East Side Health District were considered to be dependent on narcotics during a 25 month period. The age-sex-color characteristics of these individuals were used to derive population based ratios from the 1970 Census reports for the Lower East Side. See Table

ESTIMATION OF PREVALENCE OF ADDICTION NEW YORK CITY - 1971

These age-sex-color specific prevalence rates were applied to the New York City population. If New York City had the prevalence represented by MJBI's experience with Lower East Side Health District residents, there would have been a city-wide total of 104 thousand addicts.

But the prevalence of narcotic addiction in the Lower East Side is above the City-wide average and, on the other hand, MJBI was not in contact with all narcotic addicts in the Lower East Side Health District. It is possible to estimate the contrasting effects of these two factors.

Firstly, recognizing that the Lower East Side rates are higher than those of New York City, what is the estimated number of New York City addicts? The Lower East Side had the sixth highest rate of the 30 Health Districts in Koval's study. The sixth order statistic in a sample of thirty is 0.89 standard deviations from the mean (Mosteller and Rourke). Using this value of 0.89 and the geometric dispersion of 2.9 estimated earlier, we get

$$\frac{104{,}000}{(2.9)^{0.89}} = 40{,}000$$

That is, by considering the MJBI experience with the Lower East Side in terms of the ranking of the Lower East Side relative to New York City, we can estimate by order statistics that New York City would have 40,000 addicts.

However, MJBI did not see all New York City addicts during those two years. We know that MJBI was in contact with at least 25% of those first reported to the Narcotics Register in 1967; and 25% of the narcotic-related deaths in 1970. These ratios can be used to estimate the number of addicts in New York City at that time as:

$$\frac{40{,}000}{.25} = 160{,}000$$

CONCLUSIONS AND SUMMARY

There is no substitute for valid data or reliable methods in epidemiologic research. Estimates of the prevalence of narcotic addiction which are based on faulty assumptions, unstated premises, or unsubstantiated multipliers are useless. The natural history of narcotic addiction, the validity of source data, and the highly specific demographic and spatial distributions of narcotic addiction must be considered in assessing prevalence.

A log normal distribution has been identified for the prevalence of narcotic addiction in New York City and Baltimore. This log normal distribution is of major importance for analyzing the dynamics of spatial spread or diffusion.

In addition to considering trends in incidence, changes in the geometric dispersion of the distribution of prevalence by place would reflect the stage of growth or spread of epidemic disorders.

ACKNOWLEDGEMENTS

The assistance of Ms. Tanya Dubrow, and SAODAP and the reading of John Bartko, Ph.D., NIMH, is acknowledged. This investigation was supported by Research Grant Number DA-00666-01 from the National Institute of Drug Abuse.

REFERENCES

Andima H, Krug D, Bergner L et al: A prevalence estimation model of narcotics addiction in New York City. Am. J Epid 98:56-60, 1973.

Blumstein A, Sagi PC, and Wolfgang ME: Problems of estimating the number of heroin addicts in Drug Use in America: Problem in Perspective VolumeII: Social Responses to Drug Use - Appendix, 1973.

Brecher EM, Licit and Illicit Drugs Boston, Little, Brown, 1972.

Bureau of Narcotics and Dangerous Drugs. Active Narcotic Addicts. Recorded by the United States Bureau of Narcotics and Dangerous Drugs as of December, 1969. Washington, D.C., 1970.

Chambers CD: An assessment of drug use in the general population. Special Report No. 1, Drug Use in New York State New York State Narcotic Addiction Control Commission, 1971.

Chein I, Gerard DL, Lee RS, et al: The Road to H: Narcotics Delinquency and Social Policy. New York, Basic Books, 1964.

Dai B, Opium Addiction in Chicago Montclair, N.J., Paterson Smith, 1970.

DuPont RL: Profile of a heroin addiction epidemic. N Eng J Med 285:320-324, 1971.

Faris REL and Dunham HW, Mental Disorders in Urban Areas, Chicago, University of Chicago Press, 1939.

Farrell EB: Plotting experimental data on normal or log-normal probability paper. Industrial Quality Control 15: 12-15, 1958.

Greenwood JA Estimating the number of narcotic addicts. Bureau of Narcotics and Dangerous Drugs, Washington, D.C., 1971.

Jackson GW and Richman A: Risk factors in the mortality of narcotic addicts. Presented at the Annual Meeting of the Public Health Association, Minneapolis, Minnesota, 1971.

King JR Probability Charts for Decision Making New York, Industrial Press, 1971.

Koval M Opiate Use in New York City. Narcotics Addiction Control Commission and New York City Narcotics Register, 1969.

McGlothlin WH, Tabbush VC, Chambers CD et al: Alternative approaches to opiate addiction control: costs, benefits,and potential. Final Report, BNDD Contract No. J-70-33. Bureau of Narcotics and Dangerous Drugs, Washington, D.C. 1972.

Mosteller FR and Rourke REK Sturdy Statistics, Reading, Mass., Addison-Wesley, 1973.

National Commission on Marijuana and Drug Abuse. Drug Use in America: Problem in Perspective. Washington, D.C., U.S. Government Printing Office, 1973.

New York City Narcotics Register. Analysis of narcotic addiction trends through June, 1973, New York, New York.

Newmeyer JA: Estimating opiate use in San Francisco: Five methods compared. Presented at the Annual Meeting of the American Public Health Association, San Francisco, California, 1973.

Nurco D: Drug Abuse Study 1969. National Institute of Mental Health, 1970.

Richman A: Heroin addiction; social and health problems; and social environment: An area analysis of Lower Manhattan with 1970 census data. In: Proceedings of the 34th Annual Scientific Meeting, Committee on Problems of Drug Dependence, National Academy of Science, Ann Arbor, Michigan, May 22-24, 1972, pp. 963-983, 1973.

Richman A: Epidemiologic assessment of the spread of narcotic addiction. Presented at the Annual Meeting, American Public Health Association, San Francisco, California, 1973.

Richman A. The incidence and prevalence of narcotic dependency in the United States-a review of current epidemiologic information. Chapter to be published in Heroin Dependency: Social, Medical and Economic Aspects (Stimmel, B. ed.), 1974 In press.

Richman A, Feinstein M and Trigg HL Withdrawal and detoxification in New York City heroin users. In: Drug Abuse - Current concepts and research (Keup W,ed.) Springfied, Ill., Charles C. Thomas, 1972.

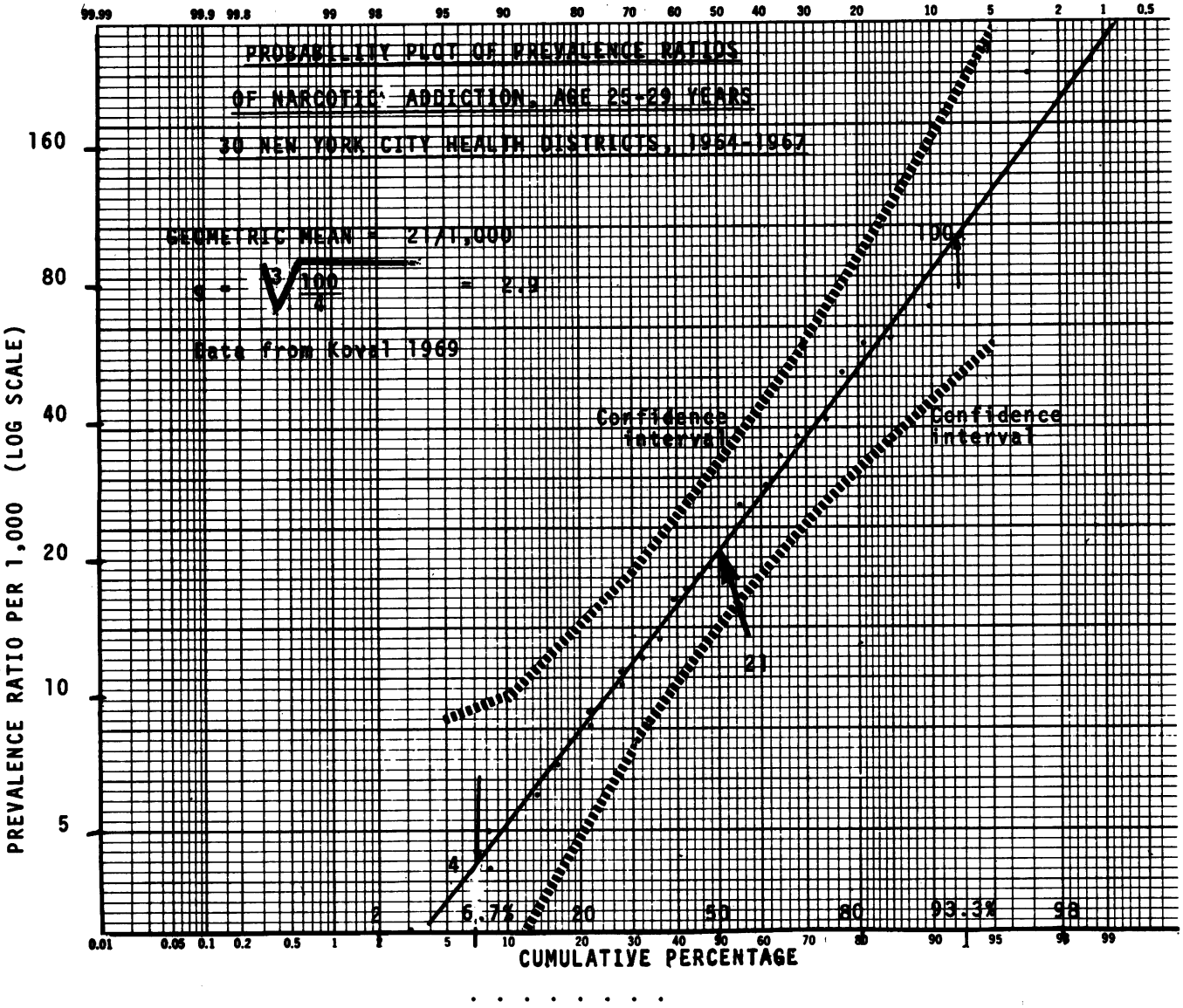
Richman A, Fishman JJ, Bergner L, et al: A narcotics case register - Some perspective on multiple reports. Social Psychiatry 6:179-185, 1971.

Robins LN, A Follow-up of Vietnam Drug Users, Special Action Office Monograph, April, 1973.

Siegel JS, Estimates of coverage of the population by sex, race, and age in the 1970 Census. U.S. Bureau of the Census, 1973.

United States Senate Committee on Government Operations, Subcommittee on Reorganization, Research and International Organizations, 93rd Congress, 1st session, Report No. 93-00. Reorganization Plan No. 2 of 1973 Establishing a Drug Enforcement Administration in the Department of Justice. U.S. Government Printing Office, September, 1973.

United States Statistical Abstract, 1969 and 1972.



M.J. BERNSTEIN INSTITUTE, BETH ISRAEL MEDICAL CENTER, NEW YORK
CONTACT WITH INDIVIDUAL NARCOTIC ADDICTS
FROM LOWER EAST SIDE HEALTH DISTRICT, MANHATTAN
TOTAL TREATED PREVALENCE, JUNE 1, 1970 - JUNE 30, 1972

RATE PER 100 POPULATION (1970 CENSUS) BY SEX, AGE GROUP AND COLOR

	MALE			FEMALE		
WHITE	15-24	25-34	35-44	15-24	25-34	35-44
Patients	860	572	134	323	132	26
Census Population	15,212	16,165	11,891	17,619	16,311	11,185
Ratio	5.7%	3.5%	1.1%	1.8%	0.8%	0.2%
BLACK						
Patients	283	295	95	102	60	23
Census Population	2,378	2,270	1,482	2,541	1,921	1,520
Ratio	11.9%	13.2%	6.4%	4.0%	3.1%	1.5%

A NOTE ON THE ANNUAL VARIABILITY OF THE CRUDE BIRTH RATE

William Seltzer and Rita S. Fand
Demographic Division, The Population Council

I. Introduction

The purpose of this paper is to investigate the extent of misinformation generated by estimating the level of the crude birth rate using data from a single year or estimating changes in this rate on the basis of data from two successive years. Briefly what we have done is assembled a few time series on the crude birth rate, computed various statistical measures that summarize aspects of the behavior of these series in time, estimated the level of random disturbance in these series using a somewhat general procedure (the variate difference technique), examined how stable these estimates of residual variability appear to be, and pointed out the implications of our results for those trying to measure or interpret short-run fertility changes.

II. Basic Data

The basic data consist of time series of annual crude birth rates in seven countries. We have used crude birth rates in this study rather than some other variable that is more closely related to the dynamics of fertility behavior for three reasons: (1) the reliance that many data users place in the crude birth rate as an important indicator of fertility behavior; (2) the ease with which time series of crude birth rates can be assembled compared with that for other fertility variables; and (3) our interest in short-run changes, where age distribution and cohort effects are minimal.

Summary statistics for the seven time series are presented in Table 1. The countries involved are: the United States (for the 61 year period 1909-69), Japan (for the 93 year period 1875-1967), England and Wales (for the 130 year period 1838-1967), Algeria (for the 80 year period 1891-1970), Sweden (for the 202 year period 1766-1967), Finland (for the same 202 year period), and Malta (for the 70 year period 1900-1969).

Together this body of data covers 834 country years of experience reflecting considerable diversity in population size and some diversity in cultural background and demographic history. However, it should be noted that the selection of countries included was arbitrary and the nature and the quality of the data used varies widely. For example, both the Algerian and the United States series are known to contain adjustments for underregistration. Despite the necessary qualifications arising from the arbitrary choice of countries and the problems of data quality, we think our results will be of interest to those trying to interpret short-run changes in fertility.

In this connection, let us point out two interesting facets of these series shown in Table 1. First, the median year to year percentage change in the crude birth rate varies from 1.0 to 5.1 among these seven countries -- in our view a rather narrow range -- and one that indicates that many annual changes in the crude birth rate will be too small to be reliably detected except by a well calibrated civil registration system. Second, note that the number of years with no change in the crude birth rate from the preceding year ranges from about 1 to 5 percent of each series. However, this is an arbitrary result depending upon the number of significant digits used in presenting the crude birth rate. In the case of Sweden if one expresses the crude birth rate as an integer rate per 1,000 (for example, 25 per 1,000) rather than using one more significant digit (such as a crude birth rate of 25.4) the number of no change years increases from 7 to 72 out of the 201 year total. The result is obvious to statisticians but has not been adequately communicated to data users.

To supplement the annual crude birth rate series from these seven countries, we also make use of ten other time series of annual crude birth rates (nine U.S. states for the 1915-1968 period, and New York City for the period 1898-1953) as well as various series of live births for England and Wales, Sweden, and New York City.

III. The Variate Difference Method

The variate difference method is a technique that permits us to estimate the level of noise of disturbance in a time series on the assumption that each term in the observed series is the sum of two unobservable components: (1) a polynomial of degree n (or less), and (2) a random disturbance element. That is, we assume in the present case that

$$b_y = f(y) + e_y \qquad [1]$$

where b_y = the (observable realized crude birth rate in some population in year y,

$f(y)$ = a polynomial in time of degree n, corresponding to the underlying trend of the crude birth rate, and

e_y = a random disturbance with E(e) = 0 and a constant variance, σ_e^2.

Given this rather general model, it is possible to estimate σ_e^2, the variance of the disturbance terms of the observed series.

Before proceeding further two points need to be made. The model is quite general. No specific polynomial is assumed nor does one even make a specific assumption about the degree of the polynomial involved. However, one must not mistake this generality with universality. Many series do not correspond to such a polynomial; for example, trigonometric or transcendental functions, functions that include independent variables in addition to time, and functions of a degree higher than n (although under a variety of circumstances each of these functions can be approximated by such a polynomial). Furthermore, each disturbance term is assumed to have no direct effect on any other disturbance term nor is the variance of the disturbance component permitted to vary over time. Because of these limitations the variate difference method has frequently led to unsatisfactory results [Kendall and Stuart, 1966; Anderson, 1971]. Fortunately, in the present situation it appears to lead to relatively stable estimates.

Making use of the fact that both the polynomial and disturbance elements of each term behave in opposite ways as successively higher order differences are taken, it is possible [Anderson, 1971 and Kendall and Stuart, 1966] to show that for a series of t terms that

$$\left(\frac{1}{t-n} \sum_{y=1}^{t-n} (\Delta^n b_y)^2 \right) \Big/ \begin{bmatrix} 2n \\ n \end{bmatrix} \qquad [2]$$

corresponds to an estimate of σ_e^2 using the differences of order n.

In practice, one avoids assuming knowledge one does not have by estimating the residual variance initially on the basis of first order differences, then using the second order differences, then the third order differences, and so on, stopping when the polynomial component has been suppressed and the variance estimate stabilizes. That is, one forms an estimate of σ_e^2 using equation 2 with n=1, then another estimate with n=2, and so on. If the model holds, the successive estimates of σ_e^2 should decline as long as n is less than or equal to the degree of the trend polynomial f(y); thereafter, when the order of differencing exceeds the degree of the trend polynomial the variance estimate should stabilize.

IV. Results

The results of applying the variate difference technique to the crude birth rate series of these seven countries are presented in Tables 2 and 3. Since few real time series can be expected to correspond exactly to the model given by equation 1, the estimates of the variance and coefficient of variation 1/ of the disturbance component of the series for each country do not stabilize as one takes higher order differences -- rather they tend slowly to drift. However, the median values for these seven series do stabilize after about the fourth or fifth difference, indicating a standard deviation of the residuals of about 3/4 of a point in the crude birth rate or a coefficient of variation of the residuals of somewhat over 2.5 percent.

The strategy of the balance of this section is to demonstrate the relative stability of these estimates of residual variability in the face of efforts to identify factors with which the residuals may be associated. The following factors are examined explicitly: first, the size of the population; second, the length of the time series; third, the time period covered by the series; fourth, the use of a time series of crude birth rates compared with one of live birth aggregates; and finally, the use of a time series of annual terms compared with one of monthly terms.

To summarize quickly the results of these analyses, we can find no clear pattern in the level of residual variability for any of the factors examined, except possibly when a monthly series is used rather than an annual one. In this case the pattern is clear -- the monthly series has a coefficient of variation for the disturbance term that is about 1 1/2 to 3 1/2 times as large as that for the corresponding annual series. This differential is maintained even if one attempts to control for series length and seasonality by looking at the year to year changes in the series for individual months. Our only qualification to this finding is that it is based on data for only one time series -- New York City -- and we find it is somewhat difficult to assert that this City is a typical place.

Some discussion of the factors that did not appear to be associated with the

estimated level of residual variability is in order. It might be conjectured that the deviations from the polynomial trend are the result of some binomial process. If this were so, then one would expect the variance and the coefficient of variation of the residuals to be larger for countries with a small population than for those with a large one. Although far from statistically significant, the shadow of such a pattern appears to be lurking in Table 2. One finds, for example, based on estimates of residual variability using differences of order five:

Midrange Population (millions)	Countries	Median CV
Under 4	Malta, Finland	3.7
4 - 24.9	Sweden, Algeria	3.2
25 - 63.9	England and Wales	2.4
64 and over	U.S., Japan	2.4

Furthermore, the correlation coefficient between population size and the estimated variance of the residuals (again based on Δ^5) is -0.49.

In order to examine whether this pattern would emerge more clearly if other possible sources of residual variation were controlled, annual crude birth rate series for nine U.S. states covering a standard 54 year period were studied. Estimates of residual variability for these series fail to confirm that the deviations from the assumed polynomial trend are associated with population size in any simple way. 2/ This finding replicates the conclusion of Hotelling and Hotelling [1931] based on a live birth time series that the variability of the residuals is substantially greater than can be accounted for by a binomial process.

Tables 2 and 3 also suggest that the longer series for the seven countries may be more variable than the shorter ones -- with respect to both the underlying birth rates (the b_y's in equation 1) and the disturbance terms (the e_y's). However, if one compares results based on standard 54 year time periods for these same seven countries with those based on the original series, a confused and partially contradictory picture emerges:

Length of Series	CV of CBR Median	Est. residual CV (using Δ^2) Median
Varied	19.6	2.7
54 years	15.3	3.6

Although this analysis standardizes for series length and nearly so for time period covered, it includes nations at various stages of demographic development. For example, fertility in Algeria remains at high, pre-industrial levels, it has been at near-replacement levels in Sweden for some years, and it has followed a generally downwards, but widely varying, path in the United States over the past few decades. In order to see if there is some association between the level of variability and where in the demographic transition a population seems to be, data for two contrasting 54 year time periods was examined for the four longest time series (that is, Sweden Finland, England and Wales, and Japan). 3/

In both absolute and relative terms the annual crude birth rates for these four countries displayed more variability in the most recent 54 year period than in the earliest 54 year period observed in each series. This is true whether one includes Japan or excludes it from consideration because of the recency of its major fertility decline. However, the pattern is less clear for estimates of the variability of the residuals. The median estimated coefficient of variation for the most recent period is somewhat greater than for the early period (for example, 4 versus 3 percent using Δ^5). On the other hand, the estimated coefficient of variation of the residuals for the Swedish series -- one of the two longest -- was higher in the earlier period than the later one.

One may justifiably speculate about the extent to which these findings are an artifact of particular features of the demographic and social history of the populations studied. Our response can only be cautionary: seven countries, nine states, and one city is an inadequate representation of worldwide diversity and our sample in time is limited to two centuries at most. Nevertheless, it seems reasonable to us to conclude that these estimates of residual variability, because of their relative stability, can be accepted as provisional bounds to our certitude pending more extensive research or more refined analysis.

1/ Throughout this paper the estimated coefficient of variation of the residual terms is calculated by dividing the square root of the estimated variance of the residual terms by the mean of the appropriate crude birth rate or live birth time series.

2/ The correlation coefficient between population size and the estimated variance of the residuals (using Δ^5) for the nine states is -0.26; the corresponding r for the seven countries and nine states combined is -0.12. If the square root of population size is used in the correlations in place of population size, the values of r become -0.48, -0.17, and -0.08, respectively, for the seven countries, the nine states, and the two combined.

3/ Because the Japanese time series covers only 93 years, there is a 15 year overlap between the two time periods used for this country.

A. Text References

Anderson, T.W., The Statistical Analysis of Time Series. New York: John Wiley and Sons, 1971.

Hotelling, H. and F. Hotelling, "Causes of Birth Rate Fluctuations." Journal of the American Statistical Association, 26, no. 174: 135-149 (June 1931).

Kendall, M.G. and A. Stuart, The Advanced Theory of Statistics, Vol. 3, London: Griffin, 1966.

B. Data Sources

Emerson, H. and H. Hughes, Populations, Births, Notifiable Diseases, and Deaths, Assembled for New York City, N.Y.: 1866-1938 from Official Records. New York: Delamar Institute of Public Health, College of Physicians and Surgeons, Columbia University, 1941.

_____, Populations, Births, Notifiable Diseases, and Deaths Assembled for New York City, N.Y.: Supplement, 1936-1953. New York: Delamar Institute of Public Health, College of Physicians and Surgeons, Columbia University, 1955.

Mitchell, B.R. and P. Deane, Abstract of British Historical Statistics. Cambridge: University Press, 1962.

Mitchell, B.R. and H.G. Jones, Second Abstract of British Historical Statistics. Cambridge: University Press, 1971.

National Center for Health Statistics, Vital Statistics of the United States, Volume I--Natality. Rockville, Maryland: National Center for Health Statistics (formerly published by the U.S. Public Health Service; issued yearly), 1937 and later years.

_____, Vital Statistics Rates in the United States: 1940-1960. Washington, D.C.: National Center for Health Statistics, 1968.

Seers, D., "A Fertility Survey in the Maltese Islands." Population Studies, 10 (1957): 211-27, 1957.

Statistika Centralbyran, Historical Statistics of Sweden, Part I: Population 1720-1967. Stockholm: Statistika Centralbyran (second edition), 1969.

Strömmer, A., Väestöllinen Muuntuminen Suomessa (Trans: The Demographic Transition in Finland), Series A:13. Tornio: The Population Research Institute, 1969.

Tabutin, D. and J. Vallin, L'état Civil en Algerie." Paper prepared for the Colloque de Démographie Africaine, Rabat, October 1972.

Taeuber, I., The Population of Japan. Princeton: Princeton Press, 1958.

U.N. Economic and Social Council, Demographic Yearbook. New York: United Nations (issued yearly), 1958 and later years.

U.S. Bureau of the Census, Birth Statistics for the Registration Area of the United States: Annual Report. Washington, D.C.: U.S. Bureau of the Census (published yearly), 1915-30.

_____, Birth, Stillbirth, and Infant Mortality Statistics. Washington, D.C.: U.S. Bureau of the Census, 1931-6.

Table 1 Summary Information about Annual Crude Birth Rate Series for Seven Specified Countries: Various Time Periods

Item	U.S.A.	Japan	England/ Wales	Algeria	Sweden	Finland	Malta
Length of series (years)	61	93	130	80	202	202	70
First year in series	1909	1875	1838	1891	1766	1766	1900
Last year in series	1969	1967	1967	1970	1967	1967	1969
Summary Values:							
Mean	23.63	28.34	25.90	40.95	27.28	32.45	31.39
Standard Deviation	3.75	5.48	7.86	5.13	6.98	7.20	6.15
Coeff. of Variation	.1587	.1935	.3034	.1253	.2557	.2219	.1960
Highest value	30.1	36.1	36.3	52.1	37.0	43.8	41.0
Lowest value	17.5	13.8	13.9	29.3	13.7	16.6	15.8
First value	30.0	25.3	30.3	39.8	33.8	41.5	39.0
Last value	17.7	19.4	17.2	48.2	15.4	16.6	15.8
Measures of Change:[1/]							
Increasing years							
Number	21	37	51	46	80	87	19
Percent of all years	35.0	40.2	39.5	58.2	39.8	43.3	27.5
Decreasing years							
Number	38	51	72	32	114	111	49
Percent of all years	63.3	55.4	55.8	40.5	56.7	55.2	71.0
No change years							
Number	1	4	6	1	7	3	1
Percent of all years	1.7	4.3	4.7	1.3	3.5	1.5	1.5
Total number of runs	25	47	62	43	99	120	23
Mean length of run (years)							
All runs	2.40	1.96	2.08	1.84	2.03	1.68	3.00
Increasing runs	1.62	1.76	1.89	2.19	1.74	1.53	1.90
Decreasing runs	3.45	2.32	2.40	1.52	2.48	1.85	4.08
Runs with no change	1.00	1.00	1.20	1.00	1.00	1.00	1.00
Mean square of successive differences	0.9948	3.7849	1.1573	8.3390	2.4403	6.0276	3.5368
Median year-to-year percentage change	2.4	2.7	1.9	5.1	2.8	3.5	2.8

1/ All measures of change are based on the n-1 series of the first differences for each country.

Table 2 Variance of Annual Crude Birth Rate (CBR) Series and Estimated Residual Variance, by Order of Difference Used, for Seven Specified Countries: Various Time Periods

Item	U.S.A.	Japan	England/ Wales	Algeria	Sweden	Finland	Malta	Median	Range
Variance of annual CBR series	14.0541	30.0682	61.7429	26.3437	48.6593	51.8714	37.8268	37.8	61.7-14.1
Variance of residual terms 1/									
Estimated from Δ^1	0.4974	1.8924	0.5786	4.1695	1.2201	3.0138	1.7684	1.8	4.2-0.5
Estimated from Δ^2	0.2612	1.2882	0.4145	3.2187	0.8378	2.7780	0.8944	0.9	3.2-0.3
Estimated from Δ^3	0.1977	1.0334	0.3837	2.7376	0.6778	2.7676	0.6951	0.7	2.8-0.2
Estimated from Δ^4	0.1631	0.8848	0.3775	2.4135	0.5961	2.7733	0.6036	0.6	2.8-0.2
Estimated from Δ^5	0.1430	0.8008	0.3791	2.1661	0.5530	2.7757	0.5453	0.6	2.8-0.1
Estimated from Δ^6	0.1308	0.7552	0.3835	1.9799	0.5299	2.7728	0.5023	0.5	2.8-0.1
Estimated from Δ^7	0.1229	0.7308	0.3891	1.8417	0.5172	2.7656	0.4725	0.5	2.8-0.1
Estimated from Δ^8	0.1176	0.7175	0.3952	1.7308	0.5101	2.7558	0.4548	0.5	2.8-0.1
Estimated from Δ^9	0.1139	0.7100	0.4015	1.6337	0.5062	2.7449	0.4444	0.5	2.7-0.1
Estimated from Δ^{10}	0.1112	0.7050	0.4080	1.5524	0.5044	2.7338	0.4337	0.5	2.7-0.1

1/ Calculated using equation 2.

Table 3 Coefficient of Variation (CV) of Annual Crude Birth Rate (CBR) Series and Estimated Residual CV, by Order of Difference Used, for Seven Specified Countries: Various Time Periods

Item	U.S.A.	Japan	England/ Wales	Algeria	Sweden	Finland	Malta	Median (percent)	Range (percent)
CV of annual CBR series	0.1587	0.1935	0.3034	0.1253	0.2557	0.2219	0.1960	19.6	30.3-12.5
CV of residual terms									
Estimated from Δ^1	0.0299	0.0485	0.0294	0.0499	0.0405	0.0535	0.0424	4.2	5.4-2.9
Estimated from Δ^2	0.0216	0.0400	0.0249	0.0438	0.0336	0.0514	0.0301	3.4	5.1-2.2
Estimated from Δ^3	0.0188	0.0359	0.0239	0.0404	0.0302	0.0513	0.0266	3.0	5.1-1.9
Estimated from Δ^4	0.0171	0.0332	0.0237	0.0379	0.0283	0.0513	0.0248	2.8	5.1-1.7
Estimated from Δ^5	0.0160	0.0316	0.0238	0.0359	0.0273	0.0513	0.0235	2.7	5.1-1.6
Estimated from Δ^6	0.0153	0.0307	0.0239	0.0344	0.0267	0.0513	0.0226	2.7	5.1-1.5
Estimated from Δ^7	0.0148	0.0302	0.0241	0.0331	0.0264	0.0512	0.0219	2.6	5.1-1.5
Estimated from Δ^8	0.0145	0.0299	0.0243	0.0321	0.0262	0.0512	0.0215	2.6	5.1-1.5
Estimated from Δ^9	0.0143	0.0297	0.0245	0.0312	0.0261	0.0510	0.0212	2.6	5.1-1.4
Estimated from Δ^{10}	0.0141	0.0296	0.0247	0.0304	0.0260	0.0509	0.0210	2.6	5.1-1.4

Source: Variance estimates in Table 2 and CBR means in Table 1.

DETERMINANTS OF NET MIGRATION AND MIGRATION EFFICIENCY IN SOUTHERN STATE ECONOMIC AREAS

William J. Serow, Julia H. Martin, Michael A. Spar
Tayloe Murphy Institute, University of Virginia

Introduction

In a recent paper by Galle and Williams (1972) migration efficiency rates for 1955 to 1960 were investigated using large Standard Metropolitan Statistical Areas (SMSAs) as the unit of analysis. The present study extends that investigation by analyzing characteristics of State Economic Areas (SEAs) within the East South Central and South Atlantic Census Divisions in relation to migration efficiency rates calculated for the 1965 to 1970 time period. One major objective of this study therefore, is to determine whether specific SEA characteristics are associated with migration efficiency; another is to locate and describe methodological problems arising when a number of the independent variavles used to explain the phenomenon of migration efficiency--or indeed any similar phenomenon--prove to be closely interrelated.

SEAs are used as the unit of analysis in this research because unlike SMSAs they permit coverage of rural and nonmetropolitan urban as well as metropolitan areas and populations while retaining relative homogeneity of social and economic characteristics. The Bureau of the Census defines two types of SEA, nonmetropolitan and metropolitan. The latter were first defined in 1950 as standard metropolitan areas with total 1940 populations of 100,000 or more. In 1960 additional metropolitan SEAs were defined due to compositional changes in some SMSAs. At that time metropolitan SEAs were defined as 1960 SMSAs with central cities of 50,000 or more and total populations of 100,000 or more. The East South Central and South Atlantic Divisions contain 143 SEAs, of which 54 are classified as metropolitan.

Following Galle and Williams (1972) and Shryock (1964), migration efficiency is defined as the quotient of net migration to an SEA (inmigrants minus outmigrants) divided by gross migration, the sum of all moves centered on the SEA (inmigrants plus outmigrants). It follows from this definition that the larger the absolute value of this quotient the more "efficient" the migration. For example, an SEA with net inmigration of 1,000 produced by 1,000 inmigrants and zero outmigrants would have a migration efficiency rate of 1.0, while the same volume of net migration produced by 5,500 inmigrants and 4,500 outmigrants would have an efficiency rate of 0.1. Despite the fact that rates of +1.0 and -1.0 denote the same degree of efficiency, these rates are qualitatively different, being achieved through different configurations of the basic rates. It is therefore appropriate to analyze separately those SEAs which are characterized by positive efficiency rates and those with negative efficiency rates since the pattern and weight of variables which are associated with either may be expected to differ.

Sources of Data and Methods of Analysis

Migration efficiency rates were calculated from data contained in Migration Between State Economic Areas (U.S. Bureau of the Census, 1972). Twenty-two independent variables reflecting general social and economic characteristics of the SEAs were derived from State Economic Areas (U.S. Bureau of the Census, 1972). To avoid artificial inflation of the explanatory power of these variables, no independent variable concerned with population change was included in the analysis. The twenty-two variables and their assigned mnemonic codes are listed in Table 1.

Three statistical techniques were used in the research. An initial zero-order correlation matrix was obtained in order to suggest the amount of multicollinearity present among the initial 22 independent variables. Factor analysis was then used to reduce multicollinearity, permitting the construction of factor scales. Both the factor scales and the unfactored variables were then used as independent variables in separate regression analyses, yielding a comparison of the explanatory power of each as well as the relative importance of the factor scales in "explaining" migration efficiency rates.

Methodology

In the first stage of the research the universe of 143 SEAs was divided into four groups, based on the criteria of metropolitan status (metropolitan or nonmetropolitan SEA) and the sign of the migration efficiency rate (positive or negative). A correlation matrix of the independent variables was then obtained for each group. Inspection of these matrices revealed that in each case approximately 20 percent of the correlations were greater than $|.50|$, a figure judged to indicate a high degree of multicollinearity.

A major problem associated with the use of interrelated independent variables in regression analysis was pointed out by Blalock (1963): as the degree of correlation between independent variables increases, the standard error of the estimates of the slope becomes quite large, decreasing the accuracy of the estimates. Consequently, it becomes difficult or impossible to use the beta coefficients as indicators of the relative importance of the explanatory variables, although R^2, the coefficient of multiple determiation, may still be used to indicate the total

TABLE 1. INDEPENDENT VARIABLES AND ASSIGNED MNEMONIC CODES

Mnemonic	Variable Description
POP17UND	Percent of SEAs population age zero to seventeen
POP65OVR	Percent of SEAs population age 65 years and over
DORMIESS	Percent of SEAs population residing in either barracks or dormitories
CHILDFAM	Percent of all families in SEA with own children age zero to five
NONWHITE	Percent of SEAs population nonwhite
URBANPOP	Percent of SEAs population classified as urban by the Bureau of the Census
BORNOUTS	Percent of SEAs population born out-of-state
MEDYEDML	Median years of education for the male population over age 25
MEDYEDFL	Median years of education for the female population over age 25
CHLDBORN	Number of children ever born per 1,000 married women age 35 to 44
LFPRMALE	Labor force participation rate, males age 16+
LFPRFALE	Labor force participation rate, females age 16+
UNMPLYED	Percent of labor force unemployed
LFPRWCU6	Labor force participation rate of women with children under age six
AGRCULTR	Percent of labor force in agricultural occupations
BLUECLAR	Percent of labor force in blue collar occupations
WHITECLR	Percent of labor force in white collar occupations
FEDERALS	Percent of labor force in government occupations
MEDFAMIN	Median family income
PERCAPIN	Per capita income
WLFREFAM	Percent of families receiving welfare payments
PVRTYFAM	Percent of families with earnings less than poverty level

amount of variance accounted for by all independent variables taken together.

In the present study, however, a second major problem, also noted by Blalock (1960: 357) would make the R^2's unreliable, since artificially large multiple correlations may be obtained if the number of variables in the estimated regression equation begins to approach the number of cases analyzed, a situation which did in fact occur when the original 143 SEAs were divided into four groups. Therefore, within each SEA group the 22 independent variables were subjected to a factor analysis, the factor matrices being rotated via varimax rotation to final solution. Varimax rotation is a type of orthogonal rotation procedure that attempts to obtain factors which are maximally independent of one another. In the ideal case each independent variable will load significantly on only one factor, with factor loadings near zero on the other factors, though in practice some multiple significant loadings usually occur. A successful factor analysis would, however, provide a solution to the two problems described above by both reducing the number of variables and maximizing their independence.

Having obtained the rotated factor matrix, it is necessary to transform the results into a form usable as a new set of independent variables, thus the construction of factor scale scores for each case. In constructing these scales, three decisions must be made: (1) which of the independent variables should be included in the scale; (2) what should be done with variables which load significantly on more than one factor; and (3) how should the variables be weighted in the scale. The scales used in the present study were constructed as follows: first, only those variables with factor loadings greater than $|.30|$ were considered significant. This significance level was chosen because the square of the factor loading represents the amount of variation in that variable explained by the factor. A factor loading of .30 is therefore equivalent to explaining about 10 percent of the variance in the variable. Second, variables loading significantly on more than one factor were eliminated from the analysis in order to maximize independence of the factor scales, reducing multicollinearity. Third, in cases where these criteria led to the elimination of all but one variable from a factor, the remaining variable was retained as an independent variable in standardized form. Fourth, where these criteria led to the elimination of all variables from a factor, the factor was discarded. The scales were then computed by multiplying the square of the factor loading of each variable selected by the value of the standardized variable and summing the results.

The results of these procedures yieled a new set of independent variables, greatly reduced in number and in the degree of their interrelationship. It remained to determine their explanatory power for migration efficiency and to compare this result with the power of the original, unfactored independent variables. As the final step, therefore, the factor scales and the original, unfactored variables were used as independent variables in separate multiple regression analyses with migration efficiency rates as the dependent variable in both cases.

Results

Factor Analysis

The results of the factor analyses of the 22 independent variables are presented in Table 2.

TABLE 2. FACTOR LOADINGS FOR TWENTY-TWO INDEPENDENT VARIABLES, BY SEA GROUP

Independent variables	Factor 1: High SES*				Factor 2: Young Families			
	NM(+)	NM(-)	M(+)	M(-)	NM(+)	NM(-)	M(+)	M(-)
POP17UND					.92	.79	.93	.93
POP65OVR					-.90	-.75	-.93	-.58
DORMIESS					.35			
CHILDFAM					.97	.91	.95	.75
NONWHITE	.38				.40	.36		
URBANPOP	.64	.63	.51	.57		.38	-.34	
BORNOUTS	.85	.66	.45					
MEDYEDML	.94	.91	.86	.92				
MEDYEDFL	.92	.91	.86	.91				
CHLDBORN	.37				.56	.67	.42	.73
LFPRMALE		.41			.77		.86	.58
LFPRFALE	-.35						.35	
UNMPLYED								
LFPRWCU6	-.35						-.31	
AGRCULTR								
BLUECLAR	-.91	-.32	-.97	-.92				
WHITECLR	.95	.77	.87	.88				
FEDERALS	.65		.60	.68	.47		.55	
MEDFAMIN	.31	.73	.56	.40	.35		.35	
PERCAPIN	.51	.76	.61	.45				
WLFREFAM		-.42				.43		
PVRTYFAM		-.65	-.41					

	Factor 3: Female Employment				Factor 4: Agricultural Employment			
	NM(+)	NM(-)	M(+)	M(-)	NM(+)	NM(-)	M(+)	M(-)
POP17UND						.41		
POP65OVR		-.40						
DORMIESS								
CHILDFAM								
NONWHITE	.32	.42		.79		.66		.32
URBANPOP					-.43		.68	.55
BORNOUTS			.68					
MEDYEDML								
MEDYEDFL								
CHLDBORN			.50			.56	-.31	
LFPRMALE		.83						
LFPRFALE	.84	.93	-.81	.91				
UNMPLYED	-.77	-.65	.63	-.41				
LFPRWCU6	.86	.87	-.71	.96				
AGRCULTR					.85	.77	-.85	-.79
BLUECLAR						-.83		
WHITECLR		-.41						
FEDERALS		-.66		.43	.30	.32		
MEDFAMIN		.39				-.49		
PERCAPIN		.34				-.41		
WLFREFAM		-.31				.56		.59
PVRTYFAM		-.37				.56		

	Factor 5: Poverty Families				Factor 6: Institutionalized Population			
	NM(+)	NM(-)	M(+)	M(-)	NM(+)	NM(-)	M(+)	M(-)
POP17UND						-.32		
POP65OVR								-.65
DORMIESS			.72		.80	.88		.85
CHILDFAM								.53
NONWHITE	.64	Factor Not Present	.75	.33			Factor Not Present	
URBANPOP					.42			
BORNOUTS			-.30			.36		.80
MEDYEDML								
MEDYEDFL			-.31					
CHLDBORN	.60		.54	.49				
LFPRMALE	-.31							.65
LFPRFALE								
UNMPLYED	.34		.57	.62				

TABLE 2. FACTOR LOADINGS FOR TWENTY-TWO INDEPENDENT VARIABLES, BY SEA GROUP (Continued)

Independent variables	Factor 5: Poverty Families				Factor 6: Institutionalized Population			
	NM(+)	NM(-)	M(+)	M(-)	NM(+)	NM(-)	M(+)	M(-)
LFPRWCU6			.31					
AGRCULTR								
BLUECLAR								
WHITECLR			-.34	-.32				
FEDERALS						.37		.31
MEDFAMIN	-.78		-.68	-.86				
PERCAPIN	-.77		-.72	-.69				
WLFREFAM	.84		.80	.55				
PVRTYFAM	.92		.88	.90				
		Factor Not Present			Factor Not Present			

* NM(+) = Nonmetropolitan SEAs, Positive Migration Efficiency Rate
NM(-) = Nonmetropolitan SEAs, Negative Migration Efficiency Rate
M(+) = Metropolitan SEAs, Positive Migration Efficiency Rate
M(-) = Metropolitan SEAs, Negative Migration Efficiency Rate

Of the six factors described in this table, four are present in all four SEA groups. We have labelled these high socioeconomic status (factor 1), young families (factor 2), female employment (factor 3) and agricultural employment (factor 4). It should be noted, however, that despite the fact that the substantive content of these factors is sufficiently similar so as to warrant identical labelling, there are differences in factor composition among the groups. The case of the variable NONWHITE on factor 2 is an example: it has moderately high positive loadings in both nonmetropolitan groups but is insignificant for both metropolitan groups. Similarly, the variables FEDERALS and MEDFAMIN have moderately high positive loadings for SEAs with positive rates of migration efficiency but are insignificant in both SEA groups characterized by negative efficiency rates.

In addition to these four "common" factors, two other factors were also generated. These are labelled poverty families (factor 5) and institutional population (factor 6) and are presented in the bottom section of Table 2. The poverty families factor is not present in the nonmetropolitan SEA group with negative efficiency rates; instead, the variables which load on this factor are found on the agricultural employment factor.

Factor 6, institutionalized population factor, is not present in the metropolitan SEA group with positive efficiency rates; variables which "belong" on this factor appear mainly on factors 2 and 5. A list of the variables used in constructing the final factor scales, together with their factor loadings, is contained in Table 3.

These factor scales, while not completely independent, display a pattern of intercorrelations of less magnitude than the independent variables. Data in Table 4 show that the highest correlation between any two factor scales is $|.38|$, whereas approximately 20 percent of the correlations among the unfactored variables were above $|.50|$. The use of factor analytic techniques has thus significantly reduced the degree of multicollinearity between the independent variables.

Multiple Regression Analysis

Tables 5 and 6 present results of the multiple regression analysis; Table 5 contains results for the 22 independent variables and Table 6 results for the factor scales. For each SEA group values of R^2, the coefficient of multiple determination, are given, as well as values of $\hat{R}^2$, an unbiased estimate of R^2. Our use of $\hat{R}^2$ is necessary due to the large number of independent variables in the regression function. Comparison of the two tables, and especially the values for $\hat{R}^2$ indicate that the explanatory power of the factor scales is much less than that of the unfactored variables. For the unfactored variables, R^2 ranges from .55 to .76, for the factor scales the range is much lower, from .04 to .48. The reduction in $\hat{R}^2$ values is difficult to explain, although we hypothesize that the situation is such that each of the independent variables explains some small proportion of the variance in the efficiency rate, independent of the contribution made by the others. In the aggregate this results in fairly high values of R^2 when all independent variables are used to estimate the regression function. The ommission of most of these variables in constructing factor scales leads directly to low values of R^2.

Despite the low values of R^2 produced when factor scales are used to fit a least squares equation, the following discussion will focus on the relationships between migration efficiency rates and the factor scales, because of the greater reliability of the beta coefficients derived from the regression of the factor scales.

For metropolitan SEAs with net inmigration, efficiency rates increase with increases in both the proportion of the labor force engaged in agricultural employment and with the SES level of the population. Conversely, efficiency is reduced as the proportion of either young families or poor

TABLE 3. FACTOR LOADINGS OF INDEPENDENT VARIABLES USED IN CONSTRUCTING FACTOR SCALES

Independent variables	Factor 1: High SES*				Factor 2: Young Families			
	NM(+)	NM(-)	M(+)	M(-)	NM(+)	NM(-)	M(+)	M(-)
POP17UND					.92		.93	.93
POP65OVR					-.90		-.93	
CHILDFAM					.97	.91	.95	
BORNOUTS	.85							
MEDYEDML	.94	.91	.86	.92				
MEDYEDFL	.92	.91		.91				
LFPRMALE							.86	
LFPRFALE	-.35						.35	
LFPRWCU6	-.35							
BLUECLAR	-.91		-.97	-.92				
WHITECLR	.95							

	Factor 3: Female Employment				Factor 4: Agricultural Employment			
	NM(+)	NM(-)	M(+)	M(-)	NM(+)	NM(-)	M(+)	M(-)
LFPRFALE		.93		.91				
UNMPLYED	Factor Dropped	-.65	Factor Dropped					
LFPRWCU6		.87		.96				
AGRCULTR					.85	.77	-.85	-.79

	Factor 5: Poverty Families				Factor 6: Institutionalized Population			
	NM(+)	NM(-)	M(+)	M(-)	NM(+)	NM(-)	M(+)	M(-)
DORMIESS			.72			.88		.85
NONWHITE			.75					
BORNOUTS		Factor Not Present			Factor Dropped		Factor Not Present	.80
UNMPLYED	.34		.57					
WLFREFAM	.84		.80					
PVRTYFAM	.92			.90				

* NM(+) = Nonmetropolitan SEAs with Positive Migration Efficiency Rates
NM(-) = Nonmetropolitan SEAs with Negative Migration Efficiency Rates
M(+) = Metropolitan SEAs with Positive Migration Efficiency Rates
M(-) = Metropolitan SEAs with Negative Migration Efficiency Rates

TABLE 4. ZERO-ORDER CORRELATION COEFFICIENTS FOR FACTOR SCALES, FOR SEA GROUPS

	M(+) Above Diagonal, NM(+) Below Diagonal			
Factor Scale	High SES	Young Families	Poverty Families	Agricultural Employment
High SES	---	.12	-.34	-.05
Young Families	.06	---	.05	-.24
Poverty Families	-.21	.05	---	-.22
Agricultural Employment	.05	-.26	.23	---

	M(-) Above Diagonal, NM(-) Below Diagonal				
	Female Employment	High SES	Agricultural Employment	Young Families	Institutionalized Population
Female Employment	---	.14	.05	.10	.24
High SES	.18	---	-.38	-.13	.35
Agricultural Employment	.20	-.29	---	.30	-.22
Young Families	.13	.05	-.03	---	-.01
Institutionalized Population	.19	.25	.03	-.01	---

families in the SEA increases. The proportion of young or poor families has a marginally greater association with migration efficiency than either of the other two factor scales. Thus, a change of one standard deviation unit in the young or poor family factors is associated with a decrease of .36 and .38 standard deviation units in the efficiency rate, compared to an increase of .24 and .27 deviation units in the efficiency rate when the high SES or agricultural employment factors increase by one unit.

It is suggested that the presence of large numbers of either poor or young families decreases migration efficiency precisely because it is these family types that are migration-prone,

TABLE 5. BETA COEFFICIENTS AND COEFFICIENTS OF MULTIPLE DETERMINATION FOR TWENTY-TWO INDEPENDENT VARIABLES, BY SEA GROUP

Independent variables	Beta Coefficients for Independent Variables			
	M(+)*	NM(+)	NM(-)@	M(-)@
POP17UND	-2.82	1.73	-1.01	1.19
POP65OVR	.06	1.81	-.51	-1.19
DORMIESS	-1.32	.34	-.42	-.31
CHILDFAM	1.29	-.22	-.49	-1.15
NONWHITE	-.24	.46	.00	.63
URBANPOP	-.92	.48	.27	-.82
BORNOUTS	**	-.08	.08	**
MEDYEDML	-1.07	.06	-.51	**
MEDYEDFL	.32	-1.09	.17	-.57
CHLDBORN	1.42	-.10	1.18	.20
LFPRMALE	1.44	-.30	.51	-.36
LFPRFALE	.19	.41	-.29	.22
UNMPLYED	.79	-.51	.36	.48
LFPRWCU6	.36	-.68	-.17	-.64
AGRCULTR	-1.01	1.37	-.46	-.44
BLUECLAR	-3.41	3.84	-.22	-2.98
WHITECLR	-2.80	3.51	-.05	-2.04
FEDERALS	**	.80	-.30	.22
MEDFAMIN	-.57	-1.19	-.89	-2.56
PERCAPIN	.72	.17	.87	1.09
WLFREFAM	-.99	-.18	-.05	-.43
R^2	.91	.85	.84	.94
$\hat{R}^2$	.55	.62	.72	.76
N	27	37	52	27

** Indicates independent variables not included in regression analysis due to insufficient tolerance level in computations.

TABLE 6. BETA COEFFICIENTS AND COEFFICIENTS OF MULTIPLE DETERMINATION FOR FACTOR SCALES, BY SEA GROUP

Factor Scales	Beta Coefficients for Factor Scales			
	M(+)*	NM(+)	M(-)@	NM(-)@
High SES	.24	.48	.14	-.29
Agricultural Employment	.27	.06	-.09	.22
Young Families	-.36	-.26	-.31	-.05
Poverty Families	-.38	-.11	.45	**
Institutionalized Population	**	**	-.25	-.28
Female Employment	**	**	-.24	-.43
R^2	.54	.34	.26	.53
$\hat{R}^2$	.46	.26	.04	.48

** No factor scale constructed.
@ Beta coefficients multiplied by -1.

* M(+) = Metropolitan SEAs with Positive Migration Efficiency Rates
NM(+) = Nonmetropolitan SEAs with Positive Migration Efficiency Rates
M(-) = Metropolitan SEAs with Negative Migration Efficiency Rates
NM(-) = Nonmetropolitan SEAs with Negative Migration Efficiency Rates

and further that these family types have a relatively high turnover rate, moving both into and out of metropolitan SEAs in large numbers. The presence of high SES families and high proportions of agricultural workers within the SEA is seen on the other hand to promote efficient migration. It is probable that metropolitan SEAs are, in fact, attracting large numbers of high SES families, and losing relatively few. The positive association with agricultural employment is less explicable, although one possibility is that agricultural employment levels act as a proxy for an areas suburbanization potential--that is, those metropolitan SEAs with high agricultural employment are those that included substantial rural areas in 1960, providing a necessary precondition for future suburbanization.

The situation just dipicted also holds, with minor exceptions, for nonmetropolitan SEAs with net inmigration. For these SEAs the high SES factor is strongly associated with migration efficiency, while the agricultural employment factor and the poor families factor have weaker associations.

The relationship between factor scales and migration efficiency rates for nonmetropolitan SEAs with net outmigration is quite complex. For this group of SEAs the agricultural employment factor was positively associated with the efficiency rate while all other beta coefficients were negative. It is especially interesting to note the negative association for the high SES factor. A possible explanation of this finding is that the outmigration stream tends to be composed of young families with relatively high SES characteristics, whereas the inmigration stream tends to be composed of older families with relatively high SES characteristics. This hypothesis is supported (but by no means proven) when we look at the age distribution of the migrants moving into and out of rural SEAs. The mean percent of outmigrants between the ages of 20 and 34 is 44 percent, compared to a mean percent of 37 in this age category for inmigrants. Individuals over age 65 constitute 4.1 percent of all outmigrants, but 5.0 percent of all inmigrants. Thus, the age distribution of the in and outmigrants is in the expected direction. A more rigorous test of this hypothesis would be to relate SES characteristics and age for both in and outmigrants, but this data is not available.

REFERENCES

Blalock, H. M., Jr., "Correlated Independent Variables: The Problem of Multicollinearity," Social Forces, 42, 2: 223-237, 1963.

_____, Social Statistics, New York, 1963.

Galle, Omer R. and Max W. Williams, "Metropolitan Migration Efficiency," Demography, 9, 4: 655-664, 1972.

Shryock, Henry S., Jr., Population Mobility Within the United States, Chicago: University of Chicago Community and Family Study Center, 1964.

U.S. Bureau of the Census, State Economic Areas, Final Report, PC(2)-10B, 1972.

_____, Migration Between State Economic Areas, Final Report PC(2)-2E, 1972.

OPTIMUM CLUSTER SIZES IN MULTI-PURPOSE SAMPLE-SURVEYS

Babubhai V. Shah and Ralph E. Folsom, Research Triangle Institute

I. INTRODUCTION

The NAEP (National Assessment of Educational Progress) in school sampling design is a six-stage design. The six stages are: (1) Region, (2) Size of Community (SOC), (3) Cycle, (4), PSU, (5) School, and (6) Student. The third stage, cycle, represents a set of pseudo-strata formed by collapsing state substrata nested within major Region x SOC categories. These psuedo-strata (cycles) were introduced to facilitate the calculation of standard errors for NAEP statistics. The first three stages are assumed fixed stratification levels and are, therefore, not subject to change. Thus, the problem of finding the optimal design is reduced to finding the configuration of PSUs, schools, and students that will provide minimum variance (maximum efficiency) at a given cost. Since the number of PSUs, schools and students are constrained by the total cost, the two independent parameters of the NAEP design are (1) the number of schools per PSU and (2) the number of students per school. The basic objective of the study has been to determine the "optimal" values of these two parameters.

To determine the optimal design, estimation of variance and cost components was required. A detailed study of the cost components for NAEP's Year-02 design (Working Paper No. 8) was available and it was decided to use the results of this study since the Year-02 design and data collection procedures closely parallel the Year-03 and 04 assessments. The relevant details are presented in Reference [3]. For the estimation of the variance components, two computer programs were adapted for the NAEP design and compared. One, using the formula by Henderson [4], which is available through the Statistical Analysis System [1] was compared to "VARCOMP", an RTI algorithm developed by Shah [7] using a formula by Seeger [6]. Computation of the variance components for several statistics indicated little numerical difference between the two techniques. However, the cost of computing variance components through SAS was approximately 25 to 50 percent higher than that by "VARCOMP". The details of the formula used in VARCOMP appear in the working paper [7].

With respect to optimality criteria, if one is interested in estimating only one statistic then the solution for the optimal design is well known [5]. However, no well-defined solution exists for the "optimal" design for many statistics. A feasible definition is developed in Section II. The results for NAEP designs for group packages are presented in Sections III and IV.

II. OPTIMALITY CRITERION

If the objective of a sample survey is to estimate only one statistic, then the usual optimality criterion is the minimum variance for the statistic at a given cost. However, it is rare, in any survey, that one is interested in only one statistic. The optimality criterion for many statistics is not quite obvious. Some possible suggestions are (a) the design that is optimal for most statistics, (b) the design that has minimum average variance at the given cost, and (c) the design with maximum average efficiency.

The average of several quantities is meaningful only if all the quantities are measured on the same scale and units. The variances of different statistics would be measured on different scales and units and, hence, the minimum average variance does not appear to be a meaningful criterion. To avoid the problem of proper scale, it is appropriate to define the efficiency of a design for a statistic. The efficiency of a design is a pure ratio with the numerator equal to the minimum variance that can be achieved by the optimal design for that statistic and the denominator is the variance of the same statistic for the given design.

The objective is to find the design with maximum average efficiency at the given cost and it would be desirable to have as small a variance of efficiencies over all statistics as possible. The trade-off between the maximum mean and minimum variance of efficiencies is not easy to define. However, in practice if the optimum is stable, then we may regard the minimum variance as a secondary criterion for selecting from several designs which are nearly optimal. It should be noted here that an ideal theoretic approach would be to consider the appropriate multi-variate distribution of many statistics.

It is not possible to obtain an explicit solution for the design with maximum average efficiency. Hence, an indirect attempt to solve the problem will be made. Moreover, the practical limitation on the sample survey design is likely to reduce the number of feasible designs to a few; for the way the cost model is defined and derived, it will be appropriate for only a few designs in the neighborhood of the current design. From a practical point of view, it will be sufficient to evaluate means and variances of efficiencies over all statistics for these few feasible designs, in order to determine the "optimal" design from among the practically feasible designs.

Let us assume there are M designs D_i, ($i = 1, 2, \ldots, M$) and N statistics Y_j, ($j = 1, 2, \ldots, N$). Let the estimates of the variance components of Y_j for PSU, school, and student be denoted by $V(j, \ell)$, ($\ell = 1, 2, 3$) respectively. The details regarding definitions and procedures for deriving variance components are given in Reference 8. If the cost function is assumed to be linear and the variable unit costs for PSU, school, and student are c_1, c_2, and c_3 respectively, then the efficiency $E(i, j)$ of the ith design which has p_i PSUs, s_i schools per PSU, and k_i students per school can be derived to be

$$E(i, j) = \frac{\text{Minimum Variance at the given cost}}{\text{Variance for the given design}},$$

where

$$\text{Numerator} = \{ \sum_{\ell=1}^{3} \sqrt{c_\ell V(j,\ell)} \}^2 \div (c_1 p_i + c_2 p_i s_i + c_3 p_i s_i k_i)$$

and

$$\text{Denominator} = \frac{V(j,1)}{p_i} + \frac{V(j,2)}{p_i s_i} + \frac{V(j,3)}{p_i s_i k_i} .$$

The expression for minimum variance at given cost is given in any sampling textbook, e.g., Murthy [5]; the expression for the variance of a given design is developed in Reference [8]. Once the efficiencies E(i, j) are computed, means and variances over j can be computed as

$$M_i \{E(i,j)\} = \frac{1}{N} \sum_{j=1}^{N} E(i,j),$$

$$V_i \{E(i,j)\} = \frac{1}{N-1} \sum_{j=1}^{N} \{E(i,j) - M_i\}^2 .$$

These means (M_i) and variances (V_i) have been used in this report for determining the optimal design.

III. RESULTS

The data used for this study consists of six packages: two group packages for each of the three age groups. For each appropriate[1/] item of these packages, forty-one (41) sets of variance components were estimated: one set associated with national p-values (the national estimated proportion of correct or acceptable answers), twenty (20) sets associated with p-values for twenty (20) domains, and twenty (20) associated Δp-values (Δp-value equals domain p-value minus the national p-value). The twenty domains that were considered are as follows: 2 by sex, 2 by race, 5 by parents' education, 4 by region, and 7 by STOC: (Size and Type of Community).

An extensive study has been carried out by Folsom and Hartwell [3] to specify a cost model for a general NAEP In-School Survey. The cost estimates were used to define 15 designs with the same costs. These designs for group packages are given, and their efficiencies are presented in Table 3.1. No efficiencies were computed for those statistics with domain p-value equal to 1 or 0. The results in Table 3.1 are based on 7578 efficiencies.

IV. COST SENSITIVITY ANALYSIS

Doubts have been raised about the accuracy of the cost estimates. It was felt that some study of possible variation in the efficiencies and "optimal" design parameters resulting from varying cost components was necessary. It should be noted here that it is not necessary to investigate the effects of such inaccuracies arising from sampling variation in the estimates of our variance components, since we are averaging over several thousand estimates. The standard deviations among estimated efficiencies are presented in Table 5.8. If the statistics were independent the standard error of a mean efficiency would be smaller than .05 percent.

Various knowledgeable persons at RTI have indicated that the errors in allocation of variable costs to PSU, school, and student are not likely to be more than 20 percent. The critical parameters that effect the optimality of designs are ratios of the variable cost components, such as (variable cost per PSU) ÷ (variable cost per school). If one of the costs is increased by 20 percent and the other is reduced by 20 percent the ratio will be altered by 1.5 or 0.66. To study the effect of the possible errors in the cost, the mean efficiencies were recomputed for all 15 designs. In each design, the number of schools per PSU and number of students per school were kept the same; however, the total number of PSUs was adjusted to make it consistent with the change in cost-ratios.

The results indicate that the mean efficiencies do not vary more than five percent under cost fluctuations within the above range. The previously obtained optimum design remains optimal or is second or third best and is within 0.5 percent in efficiency compared to the "best" one for the cost ratio. Thus, we feel that the optimal design is quite stable under cost changes.

REFERENCES

[1] Barr, Anthony James and James Howard Goodknight. A User's Guide to Statistical Analysis System. North Carolina State University, Raleigh, North Carolina 27607, 1972.

[2] Folsom, Ralph E., David L. Bayless, and B. V. Shah. Jack-knifing for Variance Components in Complex Sample Survey Designs. Presented at the American Statistical Association Meetings at Fort Collins, Colorado, August 23, 1971.

[3] Folsom, Ralph E. and T. D. Hartwell. A Study of Optimum Designs for an In-School Assessment, Working Paper No. 8, RTI Project 25U-688, 1972.

[4] Henderson, C. R. Estimation of Variance and Covariance Components, Biometrics, Vol. 9; 1953, pp. 226-252.

[5] Murthy, M. N. Sampling Theory Methods. Statistical Publishing Society: 20411 Barrackpore Trunk Road, Calcutta-35, India, 1967, p. 347.

[6] Seeger, Paul. A Method of Estimating Variance Components in Unbalanced Designs. Technometrics 12, 1970, pp. 207-218.

[7] Shah, B. V. and C. N. Dillard. VARCOMP Program for Computing Variance Components. Working Paper No. 30, RTI Project 255U-796, 1974.

[8] Shah, B. V.; Folsom, R. E. and Clayton, C. A. Efficiency Study of YEAR-03 In-School Design. RTI Final Report 255U-796-2, 1973.

Table 3.1

Mean and Standard Error of Efficiencies for 15 Cost Equivalent Designs for NAEP Study

Design	No. of PSUs	No. of Schools per PSU	No. of Students per School	Total No. of Students	Total Variable Cost	All	Standard Deviation
1	157	1	16	2,512	303764	69.9	13.8
2	179	1	12	2,148	303004	66.5	14.5
3	209	1	8	1,672	303200	59.1	15.4
4	85	2	18	3,060	300475	77.0	11.3
5	92	2	16	2,944	302952	77.2	11.6
6	99	2	14	2,772	302041	77.0	12.1
7	108	2	12	2,592	303359	76.1	13.0
8	118	2	10	2,360	302887	74.2	14.2
9	130	2	8	2,080	302224	70.7	15.6
10	60	3	18	3,240	300851	77.8	11.1
11	65	3	16	3,120	302323	78.4	11.1
12	70	3	14	2,940	300164	78.7	11.5
13	77	3	12	2,772	302225	78.5	12.3
14	85	3	10	2,550	302765	77.2	13.5
15	95	3	8	2,280	303894	74.5	15.1

THE VALIDATION OF A PETITION BY A SAMPLE SURVEY

J. J. Shuster, University of Florida

ABSTRACT

A student petition needs K signatures to be approved. A total of N signatures are obtained. This paper deals with the question: Are there more than K distinct signatures among the N? Because of time and cost factors, we shall base our inference on the number of invalid signatures in a random sample of n signatures.

1. INTRODUCTION

Consider a petition that requires K distinct signatures for validation. Suppose that N_1 signatures appear once, N_2 signatures appear twice, and so on. The number of distinct signatures, D, is:

$$D = \sum_{i=1}^{\infty} N_i \ , \qquad (1)$$

while the total number of signatures, N, is

$$N = \sum_{i=1}^{\infty} iN_i \ . \qquad (2)$$

Based on a random sample of n names, we wish to test: H_0 : D≤K against H_A: D>K, at the α×100% significance level.

The burden of "proof" is assumed to fall on the people whose interest is to make the petition pass.

As a test statistic, we shall use Y, the total number of signatures, known to be invalid, that fall in the sample. If n_i is the number of signatures appearing i times in the sample,

$$Y = \sum_{i=1}^{\infty} (i-1)n_i \ . \qquad (3)$$

We shall reject H_0 if Y falls below a specified critical value.

For any given D, the distribution on the N_i which stochastically minimizes Y can be shown to be:

$$N_1 = 2D-N \quad N_2 = N-D \quad N_3=N_4=\ldots=0, \qquad (4)$$

where D and N are given by (1) and (2).

The above is obtained by comparing iteratively Y for an arbitrary N_i scheme with Y for the scheme modified by replacing one signature of multiplicity three or more by a signature that duplicates a signature that appears once.

Since we reject for small values of Y, the significance level for the distribution of N_i given in (4), is at least as large as that of any other distribution.

We now restrict our attention to the distribution given in (4). The problem may be stated as follows: We wish to test: $H_0 : N_2 = R = N-K$ against $H_A: N_2<R$. The test statistic is

$$Y = \text{number of duplicated names in sample.} \qquad (5)$$

Des Raj (1961) studied this problem in connection with matching lists and obtained the mean and variance of Y. We shall give an expression for the exact distribution of Y, and derive a Poisson approximation. Barton (1958) has obtained Poisson limiting distributions in other "matching problems".

2. THE DISTRIBUTIONAL RESULTS

In this section we assume that (4) holds.

Let X be the number of signatures in the sample that come from the $2N_2$ duplicated signatures. By elementary counting rules, we obtain under H_0 :

$$P[Y=y|X=x] = \binom{R}{y}\binom{R-y}{x-2y}2^{x-2y}\left[\binom{2R}{x}\right]^{-1}, \qquad (6)$$

and hence, under H_0 :

$$P[Y=y] = \sum_{x=2y}^{\infty}\binom{R}{y}\binom{R-y}{x-2y}\binom{N-2R}{n-x}2^{x-2y}\left[\binom{N}{n}\right]^{-1}, \qquad (7)$$

where $\binom{a}{b}=0$ unless a,b are non-negative integers with b≤a.

Since N,n, and R will tend to be large, the exact probability given in (7) is difficult to compute directly.

The following theorem is therefore useful:

Theorem: Let N,n, and R → ∞ in such a way that for appropriate constants λ and p:

$$(a) \quad N = n^2/\lambda \ , \qquad (8)$$

$$(b) \quad R = Np = n^2p/\lambda \ . \qquad (9)$$

Then under H_0 :

$$P[Y=y] \to (\lambda p)^y e^{-\lambda p}/(y!), \qquad (10)$$

the Poisson distribution with mean λp.

Proof: Let

$$t = [2np]. \qquad (11)$$

By Stirlings approximation (see Feller (1957), page 54), and simplification, we obtain from (6)

$$P[Y=y|X=t] \sim (\lambda p)^y e^{-y} A_n B_n C_n D_n/(y!), \qquad (12)$$

where

$$A_n = \left(1 - \frac{\lambda}{n}\right)^{-t} \to e^{2\lambda p}, \tag{13}$$

$$B_n = \left(1 - \frac{2\lambda}{n} + \frac{\lambda y}{pn^2}\right)^{t} \to e^{-4\lambda p}, \tag{14}$$

$$C_n = \left(1 - \frac{y}{pn}\right)^{-t} \to e^{2y}, \tag{15}$$

and

$$D_n = A_n^{-2R/t} B_n^{-R/t} \to e^{\lambda p - y}. \tag{16}$$

Hence,

$$P[Y=y|X=t] \to (\lambda p)^y e^{-\lambda p}/(y!). \tag{17}$$

Next, from (6), we obtain for positive integers x and m:

$$\frac{P[Y=y|X=x+m]}{P[Y=y|X=x]} =$$

$$\prod_{j=0}^{m-1} \frac{(x+m-j)(R-x+y-j)}{(x+m-2y-j)(R-\frac{1}{2}(x+j))}. \tag{18}$$

Let $\delta>0$ be an arbitrary positive number and let

$$U(\delta,t) = \left[\frac{(t-\delta t^{\frac{1}{2}})(R-t+y)}{(t-\delta t^{\frac{1}{2}}-2y)(R-\frac{1}{2}(t+\delta t^{\frac{1}{2}}))}\right]^{\delta t^{\frac{1}{2}}}, \tag{19}$$

$$L(\delta,t) = \left[\frac{R-t+y-\delta t^{\frac{1}{2}}}{R-\frac{1}{2}t}\right]^{\delta t^{\frac{1}{2}}}, \tag{20}$$

and

$$M(\delta,t) = \max\{U(\delta,t), [L(\delta,t)]^{-1}\}. \tag{21}$$

Then by elementary analysis,

$$[M(\delta,t)]^{-1} \le \frac{P[Y=y|X=x]}{P[Y=y|X=t]} \le M(\delta,t), \tag{22}$$

for every integer x such that

$$|x-t| \le \delta t^{\frac{1}{2}}, \tag{23}$$

where t is given in (11).

Under the conditions of the limiting process, it is readily seen that

$$M(\delta,t) \to 1. \tag{24}$$

Let $\epsilon>0$ be arbitrary, and choose δ such that the hypergeometric random variables X satisfy (for all t given by (11)):

$$P[t-\delta t^{\frac{1}{2}} < X < t+\delta t^{\frac{1}{2}}] > 1-\epsilon. \tag{25}$$

This can be done since for all t,

$$|E(X)-t| < 1 \text{ and } Var(X) < t+1. \tag{26}$$

The value of δ may be obtained from (26) and Chebyshev's inequality.

Let A be the event defined within the probability statement (25).

By (17), (22), (23) and (24),

$$P[Y=y|X\epsilon A] \to (\lambda p)^y e^{-\lambda p}/(y!). \tag{27}$$

However,

$$(1-\epsilon)P[Y=y|X\epsilon A] \le P[Y=y] \le P[Y=y|X\epsilon A]+\epsilon. \tag{28}$$

Since ϵ is arbitrary, the Poisson limiting distribution of Y is immediate from (27) and (28).

3. FIXED FRAME ANALYSIS

Suppose that of the N signatures that have been collected, R are invalid because of duplication of valid signatures, while Q are invalid because they are "illegal" signatures. The illegal signatures in the sample can be spotted with probability one, assuming that a frame is available. (For example, a list of students, homeowners, etc. can be used to check each signature in the sample.)

If we let Z be the number of signatures within the sample and off the frame, and Y be as above, the number of frame signatures in sample, known to be invalid, we estimate T, the total number of invalid signatures by

$$\hat{T} = \frac{N(N-1)Y}{n(n-1)} + \frac{NZ}{n}. \tag{29}$$

At its stochastic minimum,

$$E(\hat{T}) = T. \tag{30}$$

$$Var(\hat{T}) =$$

$$\frac{N(N-1)R}{n(n-1)}\left\{1+\frac{(n-2)(n-3)(R-1)}{(N-2)(N-3)} - \frac{n(n-1)R}{N(N-1)}\right\} + \frac{(N-n)Q(N-Q)}{(N-1)n} - \frac{4QR(N-n)}{n(N-2)}. \tag{31}$$

To estimate $Var(\hat{T})$, we replace R by $\frac{N(N-1)Y}{n(n-1)}$ and Q by $\frac{NZ}{n}$ in (31).

4. NUMERICAL EXAMPLES

1. To illustrate the Poisson approximation:

A petition needing K=12,000 names for validation has N=14,115 signatures. Find the rejection region for type I error no larger than .20, for a sample of n=1000 names.

Solution: Our H_0 is: N_2=2115. Hence,

$$\lambda p = 10.6. \tag{32}$$

We reject the hypothesis that $N_2=2115$ in favor of the hypothesis that $N_2<2115$ if the sample reveals fewer than eight invalid signatures.

If in fact $N_2=1000$ and $N_1=12,115$, the null hypothesis would be rejected with probability .867.

One may point out that our testing method is conservative when a name appears three or more times in the sample. Although this is true, such events will be rare in practice. For example, if N=14115, with $N_1 + 2N_2 = 14095$, and $N_{10} = 2$, the probability that a name appears at least three times in the sample of n=1000, is .05.

2. To illustrate the fixed frame analysis, suppose

$$N = 16,000 \quad n = 1,000 \quad Y = 4 \quad Z = 28$$

$\hat{T} = 1473$ (estimated number of invalid signatures)

$\hat{V}(\hat{T}) = 262,456$ (estimated variance of $\hat{T}$)

$\hat{\sigma}(\hat{T}) = 512$ (estimated standard deviation of $\hat{T}$).

ACKNOWLEDGEMENT

The author wishes to thank Mr. Brian Donerly, College of Journalism, for suggesting the problem.

REFERENCES

Barton, D.E., (1958). The matching distribution, Poisson limiting forms and derived methods of approximation. J.Ro.Statist.Soc., Series B 20, 73-92.

Des Raj, (1961). On matching lists by samples, J.Am.Statist.Assoc., 56, 151-155.

Feller, W., (1968). An Introduction to Probability Theory and its Applications, Vol. 1, 3rd. ed. New York: Wiley and Sons.

A LOOK AT EQUITY

Carole Siegel, Morris Meisner, Eugene Laska,

Research Foundation for Mental Hygiene
Rockland State Hospital

I. In the Nichomachean Ethics, Aristotle considered the question of what is a fair or equitable apportionment of a resource. He considered the distribution of a resource to be equitable if the apportionment to each individual is in direct proportion to his worth to society. Many would take issue with this definition, and indeed, philosophical and ideological debate about an acceptable definition of equity continues. The issues are not simple. In a recent article in Commentary [8], the readership was challenged to define a "fair" distribution of income. Income is not the only resource for which an equitable distribution is of concern. For example, some recent work [2] has discussed equity of political representation (one man-one vote) and school integration.

Equity questions may be posed in different contexts. For example, "Is the resource distribution of a given population equitable?" Another context may pose the question "Is the resource equitably distributed among the subgroups of a population?" The authors' interests in this problem arises from concern for equity in the delivery of mental health services to population subgroups [12].

Statisticians as a rule do not attempt to make a value judgment as to what is equitable but have worked on methodology - defining curves and indices and giving their properties - to describe the dispersion of a resource over the members of a population. The principal indices which have been developed for this purpose have been based on Lorenz curves. Section II of this paper collects and mathematically organizes properties of Lorenz curves and briefly surveys some associated measures of equity. In Section III a new theorem relating the behavior of Lorenz curves to monotone hazard rates is given. Specifically, it is shown that the Lorenz curves corresponding to distribution functions admitting increasing (decreasing) hazard functions lie above (below) the Lorenz curve of the exponential distribution. Section IV introduces an equity function and an equity index which may be used to study the relative behavior of the distribution of a resource over a subpopulation with respect to the total population.

II. Lorenz curves introduced in 1905 [9] have traditionally been used by economists to describe the equity of distribution of income across a population. The curve at a fixed point u measures the percentage of total income accounted for by the uth percentile of the population ordered according to increasing income.

Let Y denote the nonnegative random variable representing the resource, G its absolutely continuous distribution function and g its density function. For those random variables Y having a finite mean E[Y], the Lorenz curve, $\ell_Y(u)$, is defined by the equation

$$\ell_Y(u) = \frac{\int_0^{G^{-1}(u)} yg(y)dy}{E[Y]} \qquad (1)$$

In this setting $\ell(u)$ may roughly be thought of as the proportion of E[Y] accounted for by the uth percentile of the distribution of Y.

By differentiating (1), we obtain an alternate definition of $\ell(u)$, [4].

$$\ell_Y(u) = \int_0^u \frac{G^{-1}(x)}{E}\,dx. \qquad (2)$$

We observe that a distribution function gives rise to a Lorenz curve, and conversely since

$$E \cdot \ell'(u) = G^{-1}(u),$$

G may be recaptured from knowledge of $\ell(u)$ and E. However more than one distribution function gives rise to the same Lorenz curve. In fact, it is easily seen that for $a > 0$,

$$\ell_{aY}(u) = \ell_Y(u).$$

These definitions have the disadvantage that the so called curve of equal distribution, $\ell(u)=u$, cannot be obtained. However, one can produce Lorenz curves arbitrarily close to the curve of equal distribution. This follows from the observation that for $a > 0$,

$$\ell_{Y+a}(u) = \frac{E[Y]\ell_Y(u) + au}{E[Y] + a}. \qquad (3)$$

Letting $a \to \infty$ the Lorenz curve approaches u.

The following properties of $\ell(u)$ are immediately obvious.

1. $\ell(0) = 0$, $\ell(1) = 1$.
2. $\ell(u)$ is monotone nondecreasing. ($\ell' > 0$)
3. $\ell(u) \leq u$.
4. $\ell(u)$ is convex.

Conversely any function ℓ satisfying these properties will be thought of as a Lorenz curve. Properties (1) and (2) imply that ℓ itself is a

distribution function whose support is the unit interval. The kth moment associated with the distribution function ℓ is given by

$$\frac{E\,[YG^k(Y)]}{E[Y]} .$$

An interesting observation is that if these moments are known for all k along with E[Y] then $\ell(u)$, and hence G(x), can be recaptured.

It is known that if G is lognormal then $\ell(u)$ is symmetric about the line $\ell = 1-u$. Kendall [7] and more recently Al-Atraqchi [1] have shown that the converse is false and have derived necessary conditions on G for the symmetry conditions to hold.

Various measures of equity of distribution have been derived on the basis of the Lorenz curve. One such measure is the fair share coefficient which searches for the total proportion of the population whose values of the resource are less than the mean value, i.e., G(E). In terms of the Lorenz curve finding the fair share coefficient is equivalent to finding the value of u at which $\ell'(u) = 1$.

A more commonly used measure, the Gini coefficient, γ, [5], is the normalized (to have maximum value of one) area between $\ell(u)$ and the curve of equal distribution,

$$\gamma = 2\int_0^1 (u-\ell(u))du.$$

For example in a recent Fortune article [10], it is noted that the Gini coefficient for the distribution of income in the U. S. has fallen over the past 35 years from .44 to .35. It is well known that γ is related to the measure of concentration also introduced by Gini, $\Delta=E[|Y_1-Y_2|]$ where Y_1 and Y_2 are independent and identically distributed. The relationship is given by the equation

$$\gamma = \frac{\Delta}{2E} \qquad (4)$$

Another measure of inequity is given by the maximum vertical distance between the curve of equal distribution and $\ell(u)$, [11]. It is easily seen that this quantity occurs at u = G(E), the fair share coefficient. This distance is referred to as the Schutz coefficient and is equal to $G(E) - \ell(G(E))$.

III. The exponential distribution gives rise to the Lorenz curve

$$\ell(u) = u + (1-u)\log(1-u).$$

γ, in general, ranges between 0 and 1, and for the exponential $\gamma = \frac{1}{2}$. In some sense this makes the exponential a dividing distribution in terms of γ. In this section we show that it is a dividing distribution for the class of distribution functions admitting monotone hazard rates. The hazard rate for a distribution G is defined as

$$h(t) = \frac{-d}{dt}\log(1-G(t)).$$

Theorem:[1] Let Y have an increasing (decreasing) hazard. Then

$$\ell_Y(u) \overset{>}{(\leq)} u + (1-u)\log(1-u). \qquad (5)$$

Proof:

The Lorenz curve for a random variable X having the exponential distribution may be rewritten in terms of the distribution function G of the random variable Y as

$$\ell_X(u) = -\int_0^{G^{-1}(u)} g(t)\log(1-G(t))dt.$$

To prove the theorem, it is equivalent to show that

$$\ell_Y(u) = \int_0^{G^{-1}(u)} \frac{tg(t)}{E}\,dt \geq$$

$$-\int_0^{G^{-1}(u)} g(t)\log(1-G(t))dt = \ell_X(u). \qquad (6)$$

According to a result from reliability theory [3], the tail of a distribution having a monotone hazard rate crosses the tail of an exponential with the same mean only once. For increasing (decreasing) hazard we have

$$\begin{aligned} 1 - G(t) &\overset{>}{(\leq)} e^{-t/E}, \quad t \leq t^* \\ 1 - G(t) &\overset{<}{(\geq)} e^{-t/E}, \quad t \geq t^* \end{aligned} \qquad (7)$$

where t^* is the point at which they cross. Thus for any u such that $G^{-1}(u) \leq t^*$, the inequality (6) is true because it is pointwise true. Consider now a u for which $G^{-1}(u) > t^*$. The inequality is no longer pointwise true. However the integral inequality holds. The proof hinges on the fact that the integrands cross once at t^* and $\ell_X(1) = \ell_Y(1) = 1$. Each side of (6) can be rewritten as a sum of integrals from 0 to t^* and t^* to $G^{-1}(u)$. The integrals from 0 to t^* may be written as 1 minus the integrals from t^* to ∞. Thus we have to show

$$-\int_{t^*}^{\infty} \frac{tg(t)}{E}\,dt + \int_{t^*}^{G^{-1}(u)} \frac{tg(t)}{E}\,dt \geq$$

$$\int_{t^*}^{\infty} g(t)\log(1-G(T))dt - \int_{t^*}^{G^{-1}(u)} g(t)\log(1-G(t))dt$$

or equivalently that

$$-\int_{G^{-1}(u)}^{\infty} \frac{tg(t)}{E} \geq \int_{G^{-1}(u)}^{\infty} g(t)\log(1-G(t))dt.$$

In this form, the inequality is pointwise true. Hence the lemma is true for all u.

The intuitive interpretation of this result is based upon inequality (7) which indicates that

for increasing hazard there is less mass in the tails of the distribution function than in the tails of the exponential. Hence there is a smaller proportion of the population accounting for high values of the resource leading to a more equitable (higher) $\ell(u)$. Similar intuition holds for the decreasing hazard case.

Corollary: Let $Y_i, i=1,2$ be independent and identically distributed with monotone hazard rate. Then

$$E[|Y_1 - Y_2|] \overset{>}{(<)} E[Y_1]$$

if the hazard is increasing (decreasing).

Proof: The proof follows from the theorem and equation (4).

If the Lorenz curve of a distribution lies above (below) that of the exponential, we cannot, however, conclude that the distribution has an increasing (decreasing) hazard rate. That is, the converse of the theorem is not true. We can easily see this from (3) since if we choose Y with an arbitrary hazard function, 'a' can be chosen so that the Lorenz curve of Y+a lies above that of the exponential. One can also construct examples for which the random variable's Lorenz curve lies below that of the exponential, while its hazard function is not decreasing (e.g., $\ell(u) = u^{2k}$, k large).

IV. If one wishes to compare distinct populations with respect to equitable apportionment of a variable such as income, computing separate Lorenz curves (and their associated Gini coefficients) for each population is a method which is reasonable. However, in the particular application of concern to the authors this approach did not seem appropriate. The application involved comparing the apportionment of a resource over the total population with its apportionment over substratas of the population. A comparison of the Lorenz curve of the total with the Lorenz curve of the substrata does not indicate whether the substrata receives its fair share with respect to the total. For example, suppose the Lorenz curve of a substrata was identical to the Lorenz curve for the total population. This indicates that the apportionment of the total resource among the entire population is identical to the apportionment of the resource restricted to the substrata. However, this gives no immediate insight into the relative position of the substrata among the entire population. Hence to describe equity of a substrata, A, with respect to the total population, an equity function $B_A(u)$ is introduced whose value at a fixed point u represents the proportion of the substrata, A, whose values are less than or equal to the uth percentile for the total population.

More formally, we consider the underlying population Ω to be decomposable into two measurable disjoint, exhaustive subsets A and $\bar{A}$. Let $Y(\omega)$, with distribution function G, be the variable of interest. Introduce the indicator function of A i.e.,

$$I(\omega) = 1, \ \omega \in A$$

$$I(\omega) = 0, \ \omega \in \bar{A}$$

for any $\omega \in \Omega$. Let

$$W(y) = P(Y \leq y \mid I=1).$$

Thus, W represents the distribution function of Y restricted to members of the population in A. The equity function for the substrata A described above is formally written as

$$B_A(u) = P(Y \leq G^{-1}(u) \mid I=1) = \tag{8}$$

$$W(G^{-1}(u)), \ 0 \leq u \leq 1.$$

(Note that since B is a distribution function it is monotonically increasing and lies between zero and one.)

If $B_A(u) \geq u$ for all u then $W(\cdot) \geq G(\cdot)$. This implies that the proportion of the substrata whose resource values are less than $y = G^{-1}(u)$ is larger than the corresponding proportion in the total population. Analogous conclusions follow for the case $B_A(u) \leq u$. Thus there is a certain inequity with respect to the resource in the substrata. If $B_A(u) = u$, for all u the resource values among the substrata are distributed exactly as over the total population. We can in fact analytically show the following lemma.

Lemma: $B_A(u) = u$ for all u if and only if Y and I are independent.

Proof: Using (8), $B(u) = u$ implies $W = G$. Hence

$$P(Y \leq y \mid I = 1) = P(Y \leq y) \ ;$$

i.e., Y and I are independent. The converse is also immediate. The case $B_A(u) \equiv u$ may be thought of as "Equity in Distribution."

If we let K(y) be the distribution function of Y restricted to the substrata $\bar{A}$ and $\alpha = P(I=1)$ then

$$G(y) = \alpha W(y) + (1-\alpha) K(y). \tag{9}$$

The relationship

$$u = \alpha B_A(u) + (1-\alpha) B_{\bar{A}}(u)$$

immediately follows from equation (9). Hence if $B_A(u) > u$ then $B_{\bar{A}}(u) \leq u$.

At the beginning of this section, an argument was given for the inappropriateness of direct examination of the Lorenz curve of the substrata as a measure of equity of the substrata with respect to the total population. However, knowledge of the curve $B_A(u)$ provides no more information than does knowledge of the Lorenz curves of the total population and the substrata's along with the respective means since this enables one to generate both G and W and hence $B_A(u) = W(G^{-1}(u)) = W(E\ell'(u))$. However since these

Lorenz curves do not indicate the relative amount of the total resource available to the substrata, A, it is advisable to plot in any application the Lorenz curve of the total population and $B_A(u)$.

In an actual application one may observe N independent observations of Y of which n are observations from the substrata A. The distribution function of Y will typically be unknown. Thus one must estimate $B_A(u)$. For points $u=i/N$, B_A may be estimated as the proportion of the n observations of the substrata whose values are below that of the ith smallest of the total. Since $\hat{B}_A(u)$ is a step function its values for other u are given by the equation

$$\hat{B}_A(u) = \hat{B}_A(i/N) \quad i/N \leq u < (i+1)/N.$$

Using the lemma one may test for "Equity in Distribution" by testing for independence of the random variables Y and I. To test $B_A(u) \equiv u$ one may use standard two sample tests since the hypothesis is equivalent to $W \equiv G$ which is equivalent to $W \equiv K$. Hence if the distribution functions over A and $\bar{A}$ are not found to differ then one cannot reject "Equity in Distribution".

In the case where B(u) crosses u, one may test for equity at a given point u_o. This is done by using Fisher's exact probability test on the 2 x 2 contingency table whose rows are the proportions of the population whose resource values are less than and more than $G^{-1}(u_o)$ and whose columns refer to the substratas A and $\bar{A}$. Thus the four entries are $B_A(u_o)$, $B_{\bar{A}}(u_o)$, $1-B_A(u_o)$, $1-B_{\bar{A}}(u_o)$.

In this context the estimator of the quantity

$$J_A(u_o) = \frac{W(G^{-1}(u_o))}{1-W(G^{-1}(u_o))} \Big/ \frac{K(G^{-1}(u_o))}{1-K(G^{-1}(u_o))},$$

has been introduced as a measure of association and large sample distribution theory is well known. The assumption of equity at the point u_o corresponds to the rows and columns being independent which corresponds alternatively to the hypothesis that $J_A(u_o) = 1$. The result $J_A(u_o)>1$ corresponds to $B_A(u_o) > u_o$, i.e., inequity towards the substrata A. Hence J_A may be viewed as an index of equity of the substrata at the point u_o. If u_o is chosen as the fair share coefficient of the total population, G(E), then J has a simple interpretation. The numerator is the relative odds of being in the under fair share group of the total population, conditional on being a member of A. The denominator is defined analogously and J is the ratio of the relative odds.

Finally we note that several more global measures of inequity suggest themselves. We propose two which correspond somewhat to the definitions given in Section II.

The first measure (analogous to the Gini coefficient) proposed is given by the equation

$$\left[\sigma = \log_2 \frac{(1-\int^{+}(B_A(u)-u)^{+}\,du)}{(1-\int^{-}(B_A(u)-u)\,du)}\right]$$

where $\int^{+}(\int^{-})$ is integrated over those values of u such that $B_A(u) \overset{>}{(\leq)} u$. The range of σ is between -1 and +1, with negative values indicating inequity towards the substrata A. The value $\sigma = 0$ does not imply equity in distribution although it does in some sense imply that inequities toward A are compensated by inequities toward $\bar{A}$, albeit at different values in the range of the resource values.

Analogous to the Schutz coefficient, we can define the maximum vertical distance between the forty five degree line and $B_A(u)$. This maximum may occur at several points u. In a given application where N independent observations of the resource are made, quantity $D_N = \sup_u |\hat{B}_A(u)-u|$ where $\hat{B}_A(u)$ is the step function whose estimates at the points $u = i/N$ are described in Section III. The asymptotic distribution of D_N (which is the Kolmogov-Smirnov statistic) is well known. Since the equation $B_A(u) = u$ is equivalent to the hypothesis $G = W$, this statistic may also be used as a test of the two sample problem. Hajek [6] has commented on the relationship between rank order tests and the Kolmogorov-Smirnov test. Further for any given confidence level α, the quantity $d(\alpha)$ obtained from the asymptotic distribution of D_N such that

$$\lim_{N\to\infty} P\{D_N \geq d(\alpha)\} = \alpha$$

may be used to form an α level confidence band for $B_A(u)$, namely

$$(\hat{B}_A(u) - d(\alpha),\ \hat{B}_A(u) + d(\alpha)) .$$

1. A referee has pointed out that this result is a corollary of theorem 7 of The Estimation of the Lorenz Curve and Gini Index by J.Gastwirth appearing in The Review of Economics and Statistics, 2/72.

References

1. Al-Atraqchi, M.A. "The Symmetry of the Lorenz Curve Generated by the Log Normal Distribution Revisited", The New York Statistician, Vol. 24, No. 1, 1972, pp. 4-5.

2. Alker, Hayward, R., Jr. "Measuring Socio-Political Inequity", Tanur et al, Editors, "Statistics" A Guide to the Unknown", Holden-Day, 1972, pp. 343-351.

3. Barlow, Richard E. and Proschan, Frank "Mathematical Theory of Reliability", John Wiley and Sons, 1965, pp. 28-29.

4. Gastwirth, Joseph, L. "A General Definition of the Lorenz Curve", Econometrika, Vol. 39, No. 6, pp. 1037-1039.

5. Gini, Corrado "Measurement of Inequality of Incomes", Econ. Journal, March, 1921, pp. 124-126.

6. Hajek, J. and Sydak, Z. "Theory of Rank Tests", Academic Press, 1967, p. 90.

7. Kendall, M. G. "Discussion on paper by Hart and Prais", Journal of Royal Statistical Society, Series A, Vol. 119, Part 2, pp. 184-185.

8. Kristol, Irving, "About Equality", Commentary, Vol. 54, No. 5, November, 1972, pp. 41-47.

9. Lorenz, M. C. "Methods of Measuring the Concentration of Wealth", Publication of A.S.A., Vol. 9, 1905, pp. 209-219.

10. Rose, Sanford "The Truth about Income Inequality in the U.S.", Fortune, December, 1972.

11. Schutz, Robert "On the Measurement of Income Inequality", Am. Econ. Review, Vol. 41, May 1951, pp. 107-122.

12. The National Institute of Mental Health, Contract No. HSM-42-72-212.

SAMPLING THEORY AND ORDINAL MEASURES OF ASSOCIATION: A SOCIOLOGICAL PERSPECTIVE AND A CRITIQUE

B. Krishna Singh, Virginia Commonwealth University

Despite the emergence and reemergence of various ordinal measures of association, which have been extensively used in social research, little attention has been paid to explicit and implicit sampling theories behind such measures of association. This paper examines the nature of sampling theories associated with selected ordinal measures of association. It is assumed that if we have to continue using ordinal measures of association, it is better to use those measures which have known distributions as compared to those which have no known sampling distributions. Such an assumption seems consistent with the notion that theory construction is an implicit, if not explicit, aim of social research. In order to construct social theories (with ordinal measures of association), it is imperative that we deal with those measures of ordinal association which can provide some basis for significance testing rather than those which provide no grounds for such testing. If nothing else, measures with known sampling distributions are at least better for inferences concerning monotonic functions. As is usually the case, any inference is partly derived from statistical significance and mostly derived from substantive theoretical model to be used.

The Notion of Ordinality in Social Research

Social researchers have found that ordinal measures of association are not only appropriate in a variety of social research situations, they also involve fewer assumptions which have to be met. Even though somewhat less elegant that interval levels of relationships, ordinal associations provide appropriate and pragmatic measures which can be used meaningfully in many social research situations. The general notion of ordinality in associational terms implies that if there are two ordered politomies, they can be meaningfully associated provided there are some logical and empirical bases to assume such an association (c.f. Davis, 1967, Kim, 1971). Such a notion of association is based on the premise of transitivity asserting that if x is greater than y, and y is greater than z, then it follows that x is greater than z. The degree of greaterness is either immaterial or irrelevant or both (c.f. Coleman, 1964) in a truely ordinal sense. The notion of not knowing (and for that matter not caring to know) the magnitude of differences is usually based on the fact that sometimes it is either difficult or sometimes it is largely irrelevant to know these differences. The latter part of the above statement is a decision which has to be reached by the reseacher.

There are some prevailing controversies about the use of ordinal variables in general and use of ordinal measures of association in particular. The first controversy centers around the notion that since the basic intent of social research is to be as scientific as possible (and one way to achieve such an illusive goal is through the process of mathematization), the utility of ordinal variables is obviously limited since they cannot provide point estimations (c.f. Labovitz, 1970; Wilson, 1971). In fact, many suggestions have been made where ordinal variables should be treated as interval variables. The support for such an argument is based upon the assertion that better estimations become possible through such a treatment even though there may be some assumption violations. In addition, the argument also derives its support from the notion of statistical robustness if interval levels of measurement can be assumed.

The second controversy seems more germaine to this paper. It is based on a three dimensional view of interval measures of relationships, ordinal measures of association, and significance testing. Wilson (1971), for example, has seriously questioned the use of ordinal measures of association in the development and modifications of explanatory theories formulated in mathematical, axiomatic or deductive forms. In fact, he along with others advocate use of interval levels of measurement for constructing causal models. In addition, critiques of ordinal measures argue that since many of the ordinal measures have no known sampling distributions, there utility is of somewhat of a dubious nature in constructing models of any kind of model.

At the same time, a counterargument for such a position is proposed by stating that it is neither needed nor necessary to formulate social theories in axiomatic or deductive forms. In fact, a whole school within sociology and social sciences has recently advanced almost no quantification of social data much less go even as far as to get into the argument of ordinal versus interval levels of measurement and associations and relationships. Perhaps they may have a point but such a discussion is beyond the realm of this paper.

The question which is of immediate concern is whether some of our variables make more sense as ordinal variables or whether they should be treated as interval variables. It may be pointed out that it is usually not enough on the researchers' part to assume that his or her data meet the criterion of interval level of measurement but in fact it is necessary to demonstrate the nature if reflexiveness and transitivity in precise and accurate terms. Leaving aside the question of scaling technique, there are some questions as to whether there are indeed true interval levels of measurement in social variables and whether those variables (which for argument sake can be considered at interval levels of measurement) make more conceptual sense at ordinal variables. It is one thing to ask for more precise and accurate levels of measurement, it is another thing whether such measurements do in fact exist in a real sense. At the same time, it could be also argued that in fact we have only two levels of measurement i.e. nominal and ratio, and the distinctions between ordinal and interval levels of measurement are not very meaningful.

The most common form of use of ordinal associations in making predictions about the dependent variable y, from an independent

variable x, rests on the assumption that the independent variable predict the order properties of the dependent variable. Such order properties can be tied, concordant or discordant. The general model of such predictions is known as PRE or "Proportionate Reduction in Error" of the form:

$$PRE = \frac{E(1) - E(2)}{E(1)}$$

Distributions of Ordinal Measures of Associations:

In this section, a brief examination of five measures of association will be made. These measures are (1) tau, (2) gamma, (3) dyx (4) ordinal consensus (5) Robinson's A. It can be safely assumed that ordered polytomies with two or more categories conform to multinomial distributions. The basic postulate underlying the binomial distribution can be generalized to situations with two or more classes. Such a generalization follows the rule that: "if there are C classes, mutually exclusive and exhaustive, and with probability of $p_1, p_2, \ldots p_n$. If N observations are made independently and at random, then the probability that exactly n_1 will be of kind 1, n_2 of kind 2 and n_c of kind c is given by:

$$\frac{N!}{n_1!\, n_2! \ldots n_c!} (p_1)^{n_1} (p_2)^{n_2} \ldots (p_c)^{n_c}$$

Given any discrete probability distribution, one can easily workout the probability of all possible samples. Then, in terms of these probabilities of the disagreement of sample and theoretical distribution can be evaluated. However, the development of distributions for ordinal measures of association has been hampered by the sheer number of computations involved in multinomial distributions (c.f. Hays, 1963).

As has been suggested by Goodman and Kruskal (1963), any cross-classification follows either a multinomial distribution or a hypergeometric distribution. If the sampling is with replacement, the distribution is multinomial and if the sampling is without replacement, the resulting distribution is a hypergeometric distribution. Although, with large samples, the resulting probabilities make very little difference, the differences are hightened when we deal with samples of small sizes.

Kruskal (1958:844) has suggested that a measure of association should contain (1) simplicity of interpretation, (2) reasonable sensitivity to form of distribution, and (3) relative simplicity of sampling theory. It can be said with reasonable degree of certainty that most measures of association meet the first criterion. However, the problems arise at the last two criteria. It would appear that only measures which can meet all three criteria are (1) Gamma, and (2) Morris' adaptation of Sommer's Dyx for the bivariate case.

Ordinal Measures of Association and Theory Building in Social Research.

Up to this point, the effort has been directed toward assessing some of the problems we encounter when we use ordinal measures of association and especially those measures which have no known sampling distributions. It has been suggested that we should attempt to use only those measures which have known distributions for the simple fact that PRE interpretations are much more clearer in such measures of association.

We will operate with the premise that often times it is neither reasonable nor necessary to treat ordinal variables as interval variables and for that matter the practice can be of questionable value under most circumstances. It is not being suggested that we abandon our approach toward mathematization but it should be kept in mind that a faith blinded by trust in mathematical jargon rather than the logic of mathematics is no panacea for constructing causal models. It might be added that the process of mathematization and for that matter use of higher levels of statistical techniques is not only commendable but a necessary first step toward our eventual goal of theory construction from axiomatic and deductive perspectives. But we must be aware of what our inputs are in constructing such models.

The argument that there are certain social variables which will never meet the assumption of interval levels of measurement and consequently be never subject to higher levels of statistical analysis is a valid if not necessarily a comfortable argument. As is clear from various methodological discussions, making mere assumption about interval levels of measurement and arriving at point estimations do not necessarily mean that we are dealing with any more real phenomena than using ordinal measures of associations for constructing monotone increasing or decreasing models.

In using ordinal measures of association in theory construction, there are certain rules that we can abide by. The first rule concerns the nature of explanation on the basis of ordinal measures of association. The term "explanation" implies the nature of description which can be offered from a variable about another variable. It would appear that PRE interpretations are not only consistent but are actually well within the criterion of what we generally mean by scientific explanation.

The second rule concerns the notion of prediction. As has been noted by Wilson (1971) and Kim (1971), most measures of ordinal association are totally predictive at least at the bivariate level. However, models of ordinal predictiveness usually fail to minimize errors or expected errors of predictions. Thus, we need to be careful about how such predictions are used in theory construction. It may be noted that most of the ordinal measures of association tend to predict in the direction of concordance (depending upon the nature of scaling). One should be careful in terms of making predictions from ordinal measures of association in that they do not offer what we generally refer to as unbiased estimators.

The third rule is concerned with the nature of substantive theory building which is to be pursued. It is a cardinal fact that measures of association or relationship exist as aids rather than determinants of theory construction. When one uses ordinal measures of association for theory construction, one is simply offering building blocks for a theoretical framework which can be tested in more precise and accurate terms provided measurement problems can be

resolved. In other words, the level of theoretical abstractions that we will be able to deal with are going to be somewhat less precise but necessary steps toward a greater formalization of theoretical model.

Discussion

The major concern of this paper is that if one must use ordinal measures of association in social research, one should try to use only those measures which have known distributions. The suggestion is based on the premise that such measures provide a clearer interpretation as PRE measures and provide some notion concern-the power of statistics. In addition, with known distributions, it also becomes possible to do significance testing.

It is further suggested that from a theory construction point of view, ordinal measures of association can be as effective as any other measure if they are used carefully in conjunction with substantive theoretical efforts. While the use of ordinal measures of association may not be totally consistent with our professed goals of deriving axiomatic and deductive models, it would be imperative that we keep in mind the utility of such measures till we have arrived at ways and means of defining social variables in ratio or interval terms.

What perhaps is needed is an effort toward the development of ordinal measures of association which can take into account the nature of prediction usually associated with interval levels of relationships without loosing the three benefits proposed by Kruskal (1958).

References

Davis, J. A.
1967 "A Partial Coefficient for Goodman and Kruskal's gamma." Journal of the American Statistical Association. 62(March):189-193.

Goodman, Leo A. and William H. Kruskal
1954 "Measures of Association for Cross Classification." Journal of the American Statistical Association. 49(December):732-763.

Goodman, Leo A. and William H. Kruskal
1963 "Measures of Association for Cross Classifications. III Approximate Sample Theory." Journal of the American Statistical Association. 58(March): 310-364.

Hays, William
1963 Statistics: New York:Holt, Rinehart and Winston.

Kendall, Maurice G.
1959 Rank Correlation Methods. 2nd ed. London:Griffin.

Kim, Jae-on
1971 "Predictive Measures of Ordinal Association." American Journal of Sociology. 76(March):891-907.

Kruskal, William H.
1958 "Ordinal Measures of Association." Journal of the American Statistical Association. 53(December):814-861.

Labovitz, S.
1970 "The Assignment of Numbers to Rank Order Categories." American Sociological Review. 35(June):515-524.

Leik, Robert K.
1969 "A Measure of Ordinal Consensus." Pacific Sociological Review. 9(fall): 85-90.

Morris, Raymond N.
1970 "Multiple Correlation and Ordinally Scaled Data." Social Forces. 48(March) :299-311.

Robinson, W. J.
1957 "The Statistical Measurement of Agreement." American Sociological Review. 22(February):17-25.

Sommer, Robert H.
1962 "A New Asymmetric Measure of Association for Ordinal Variables." American Sociological Review. 27(December):799-811.

Wilson, Thomas P.
1971 "Critique of Ordinal Variables." Social Forces. 49(March):432-444.

UNDERREPORTING OF BIRTHS AND DEATHS IN HOUSEHOLD SURVEYS OF POPULATION CHANGE

Monroe G. Sirken and Patricia N. Royston
National Center for Health Statistics

A. Introduction

In this paper we report the results of a survey experiment that was conducted to investigate the effect of different counting rules on the completeness of enumeration of births and deaths in single retrospective surveys of population change. In this type of survey, households report retrospectively those births and deaths that occurred during a prior calendar period, also referred to as a reference period. Counting rules in the single retrospective survey specify the conditions for linking persons who experienced the vital events during the reference period to the housing units where they are eligible to be counted in the survey.

In the survey experiment, we investigated the following counting rules for enumerating births and deaths:

Rule	Statement of Rule for Enumerating Deaths
1	The death is enumerated at the decedent's former residence.
2	The death is enumerated at a housing unit adjacent to the decedent's former residence.
3	The death is enumerated at the residences of surviving siblings exclusive of the decedent's former residence.
4	The death is enumerated at the residences of surviving children exclusive of the decedent's former residence.

Rule	Statement of Rule for Enumerating Births
1	The birth is enumerated at the mother's residence at the time of birth.
2	The birth is enumerated at the mother's residence at the time of the survey.
3	The birth is enumerated at the housing unit adjacent to the mother's residence at the time of birth.
4	The birth is enumerated at the residences of maternal siblings exclusive of the mother's residence.
5	The birth is enumerated at the residence of maternal grandparents exclusive of the mother's residence.

Any one of the above counting rules or combinations of several of them might be adopted in the survey to enumerate births or deaths. Generally, surveys of population change adopt counting rules which have the property of linking every vital event that occurred during the reference period to one and only one housing unit where it would be eligible to be enumerated in the survey. Clearly, it is desirable to adopt a rule that links every event to a household since the unlinked events would be missed in the survey. However, counting rules need not be restricted to those which uniquely link every event to one and only one housing unit, since unbiased estimates have been developed [Sirken, 1970a] for counting rules which link vital events to multiple housing units.

In designing the survey, the optimum counting rule strategy is the selection of counting rules which minimize the mean square error of the survey estimates. In prior reports, we [Sirken, 1970b, 1972] have investigated the effect of alternative counting rules on sampling errors of survey statistics. In this paper, we compare the bias due to underenumeration of births and deaths associated with different counting rules in single retrospective surveys of population change. This is an extension and refinement of a prior paper [Sirken and Royston, 1970] in which we compared the counting rule bias of different rules for enumerating White deaths in single retrospective surveys.

B. Survey Experiment

The experiment was based on a sample of 284 noninstitutionalized deaths and 285 legitimate births that occurred in Los Angeles during the four month period July-October 1969. The sample events, approximately equally divided between Whites and Blacks, were selected from the vital record files of the Los Angeles County Department of Health.

The household survey experiment was conducted during the three month period January-March 1970. Thus, from three to nine months elapsed between the dates of occurrence of the sample events and the dates the households were contacted in the survey. The survey was fielded in two stages. First, interviews were conducted at the places of residence listed on the vital records of the sample events. We will henceforth refer to the residence address on the vital record as the key address. On the death record, it represents the usual place of residence of the decedent at the time of his death and on the birth record it represents the usual place of residence of the mother when the baby was born. Second, interviews were conducted in Los Angeles County at the housing units of the surviving siblings and children of the sample decedents and at the housing units of the maternal aunts, uncles and grandparents of the sample births. These interviews were limited, however, to relatives who were identified and whose addresses were ascertained in interviews that were completed at key addresses. Possibly other relatives existed who were not identified at key addresses and one might speculate that they would be less likely to

report the events than the relatives who were identified at key addresses.

The household respondent at the key address was asked to identify births and deaths which had occurred at that address during the prior 12 month period. The respondent at an address of the decedent's surviving siblings and children was asked to identify deaths that occurred during the prior 12 month period to any siblings and parents of persons living in the household at the time of the interview. And the respondent at an address of the mother's siblings and parents was asked to identify the births during the prior 12 month period of nieces, nephews, and grandchildren of persons living in the household at the time of the interview.

A proxy respondent rule was used at key households and at the households of relatives, namely any adult in the household was an eligible proxy respondent for himself and for all other household members. Perhaps the completeness of enumeration of vital events, particularly at households of relatives, would have been greater had the experiment used a self-respondent rule in which every adult responded for himself.

C. Findings for Deaths

The findings of the experiment with respect to deaths are summarized in Table 1 separately for Whites and Blacks. Counting rules 2, 3, and 4 are based on small samples making the analysis tenuous and difficult. The findings are presented primarily as illustrative rather than as firm estimates. For each counting rule, the proportion of deaths missed in the experiment was greater for Blacks than for Whites. For Blacks, the proportion missed was uniformly high, about 40-45 percent for each counting rule. For Whites, a larger proportion was missed at key housing units and at residences of surviving siblings than at the residences of surviving children and residences of neighbors. Migration of the decedent's household between the date of his death and the survey date contributed substantially to the number missed at the key housing units.

There were two distinctly different reasons why vital events were not enumerated in the experiment. They were missed either because contact was not established with the household or because the events were not reported in interviews that were conducted. For both Blacks and Whites, more of the deaths were missed because events were not reported in conducted interviews than because interviews were not conducted (including not-at-homes, and vacant housing units). We had a problem in deciding whether to classify refusals as "interview conducted" or "interview not conducted." We decided it made more sense to consider them as conducted interviews since contact was established with the household. In the earlier report [Sirken and Royston, 1970], missed deaths were subdivided into two groups depending on whether or not the interview was completed, and in that report refusals were classified as interviews that were not completed. Consequently, the statistics presented in the two reports may appear to be somewhat inconsistent.

Since there is variation among the counting rules in the proportion of interviews that were conducted, we have estimated ρ for each counting rule. This parameter represents the proportion of events that were missed in the subsample of households where interviews were conducted. Estimates of ρ separately for White and Black deaths are presented in Table 2. For every counting rule, the estimates of ρ are uniformly lower for Whites than for Blacks. For both Whites and Blacks, however, the estimate of ρ is smallest for counting rule 4. That is, fewer White and Black deaths were missed at the housing units of surviving children than at the housing units linked to deaths by counting rules 1, 2, and 3.

D. Findings for Births

The findings for births are summarized in Table 3. For both Blacks and Whites less than 5 percent of the births were missed by rule 2, which links births to the mother's survey residence. For each of the four other rules, the proportion of Black births missed in the survey substantially exceeds the proportion of missed White births. Exclusive of rule 2, the proportion of White births than were missed ranged from 10 percent for births linked to grandparents to 14 percent for births linked to neighbors and for Blacks the proportion of deaths missed ranged from 34 percent for births linked to key addresses to about 60 percent for the other rules.

Estimates of ρ for births are presented in Table 4. For both Whites and Blacks rule 2 linking births to the mother's survey residence stands out prominently as the best rule. Estimates of ρ for the other rules ranged for Whites from about 10 percent to 20 percent and for Blacks from about 18 percent to about 60 percent. It may be of interest to note that the four White births missed at the household of grandparents were due to refusals.

E. Summary

We have presented findings based on a small survey experiment that was conducted to determine the effect of counting rules on the completeness of enumeration of births and deaths in single retrospective surveys of population change. Although the experiment was based on small samples of vital events and was subject to other design limitations, it appears (1) that counting rules have a substantial effect on the extent of underenumeration of births and deaths in household surveys and (2) that enumeration of births and deaths was more complete for Whites than for Blacks regardless of the counting rule used in the experiment. Also, we believe that the overall level of enumeration completeness in the survey experiment could be substantially improved by increases in survey resources and improvement in the survey methodology.

Table 1. Percent of adult deaths by color that were missed, by type of counting rule tested in the survey experiment

Color	Counting Rule[1] for Enumerating Deaths			
	1	2	3	4
WHITE				
Number of deaths	139	25	15	29
Total percent	100	100	100	100
Deaths reported	65	76	67	83
Deaths missed	35	24	33	17
Interview conducted	25	20	27	14
Interview not conducted	10	4	7	3
BLACK				
Number of deaths	145	30	12	26
Total percent	100	100	100	100
Deaths reported	55	60	58	58
Deaths missed	45	40	42	42
Interview conducted	28	33	42	12
Interview not conducted	17	7	0	31

[1]See text for definition of counting rules.

Table 2. Estimates of ρ for Deaths by Counting Rule and Color

Color	Counting Rule[1] for Enumerating Deaths			
	1	2	3	4
WHITE				
Number of interviews	125	24	14	28
ρ	.28	.21	.29	.14
BLACK				
Number of interviews	120	28	12	18
ρ	.33	.34	.42	.17

[1]See text for definition of counting rules.

ρ = The proportion of vital events that were not enumerated in households where the interviews were conducted.

Table 3. Percent of legitimate births by color that were missed, by type of counting rule tested in the survey experiment

Color	Counting Rule[1] for Enumerating Births				
	1	2	3	4	5
WHITE					
Number of births	148	119	29	47	42
Total percent	100	100	100	100	100
Births reported	81	99	76	85	90
Births missed	19	1	24	15	10
Interview conducted	15	1	17	11	10
Interview not conducted	4	0	7	4	0
BLACK					
Number of births	137	90	29	29	20
Total percent	100	100	100	100	100
Births reported	66	97	34	34	45
Births missed	34	3	66	66	55
Interview conducted	15	3	52	55	40
Interview not conducted	20	0	14	10	15

[1]See text for definition of counting rules.

Table 4. Estimates of ρ for Births by Counting Rule and Color

Color	Counting Rule[1] for Enumerating Births				
	1	2	3	4	5
WHITE					
Number of interviews	142	119	27	45	42
ρ	.15	.01	.19	.11	.10
BLACK					
Number of interviews	110	90	25	26	17
ρ	.18	.03	.60	.62	.47

[1]See text for definitions of counting rules.

ρ = The proportion of vital events that were not enumerated in households where the interviews were conducted.

References

Sirken, Monroe G. [1970a] "Survey strategies for estimating rare health attributes." Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. III, 135-44.

Sirken, Monroe G. [1970b] "Household surveys with multiplicity." Journal of the American Statistical Association 65, 257-66.

Sirken, Monroe G. and Patricia N. Royston [1970] "Reasons deaths are missed in household surveys of population change." Social Statistics Section Proceedings of the American Statistical Association, 361-64.

Sirken, Monroe G. [1972] "Stratified sample surveys with multiplicity." Journal of the American Statistical Association 67, 224-27.

TREND SURFACE ANALYSIS AS A TOOL FOR HANDLING SPATIALLY-DEFINED NON-COINCIDENT DATA[1]

David L. Sjoquist, Georgia State University
Larry D. Schroeder, Georgia State University
Frank P. Jozsa, Jr., Georgia State University

Smoothing of economic time series data has a long history within economics. It is not uncommon to find economic practitioners estimating quarterly data from observed annual data or even estimating annual data from decenial Census data. An analogous data problem exists when analyzing social behavior of groups within a particular geographically-defined area. It is common for different (social) data sets to be defined for alternative, non-coincident geographical areas, e.g. Census tracts and voting precincts. Aggregation requires loss of information and degrees of freedom. Further, if no spatial boundaries are coincident, a smoothing or interpolation method must be used to make the data sets compatible.

Several analytical methods exist for smoothing areal data including Fourier analysis, filtering theory, or a combination of Fourier and gravity analysis.[2] However, in this paper we concentrate on a single regression technique for analyzing spatial data -- trend surface analysis (TSA) and explore its use in the preparation of data. Although emphasis is placed on TSA, the explicit objective of this paper is to compare the accuracy of several alternative methods in predicting spatially-defined, unobserved data.

After reviewing the underpinnings of TSA, we explain, in Section 2, two simple prediction methods which are used as a basis of comparison in the empirical portion of the paper. In Section 3 the prediction experiment is described with the final section containing the experimental results and the conclusions reached.

Section 1 - Trend Surface Analysis

The concept of TSA may be most easily explained in terms of a sample of data observed in either a random or regular spatial pattern. Assuming that some variable of interest, z, is measured at each of a number of geographical points, then each such point can be assigned a unique (x,y) coordinate relative to some common origin.

It is possible that the points, rather than being discrete points from a random sample, are summarizations of the level of the variable in a well-defined area immediately surrounding the (x,y) position on the map. One can thus draw a three dimensional solid with the height of the solid representing the aggregate or summary measure of a variable z within the prescribed area. For most physical and much social data it is also reasonable to suspect that the level of z for points near the boundary of any one areal unit would be, in part, associated with the level of z in the adjacent unit. Thus, one would suspect that one solid would blend into the surrounding solids so that the surface over the entire space is relatively smooth. Summarizing the surface over the space is, then, the goal of any surface analysis, where the method used in TSA is least squares regression.

Of course, least squares techniques assume specification of some form of the regression function. Unfortunately there is no theoretical justification for any particular functional form; therefore, a simple and pragmatic choice, followed throughout the remainder of the analysis, is to limit the analysis to four functional forms which creat 1st through 4th degree polynomial surfaces.

For fairly small areas and many variables it is reasonable to supect that the values of z in adjacent areas are positively correlated; however, this need not be the case. If there is no observed correlation in contiguous areas, it is reasonable to suspect that regression techniques using polynomial surfaces will not explain much of the geographical variation in z. Thus, before turning to alternative prediction methods, it is appropriate to consider one measure of the strength of areal association of a variable -- the "contiguity ratio", c, developed by Geary (1954)[3] with

$$c = \left[\frac{(n-1)}{2K_1}\right] \left[\frac{\Sigma'(z_t - z'_t)^2}{\Sigma_t (z_t - \bar{z})^2}\right]$$

where n = total number of areal units
t = any one unit
z = the variable being analyzed
$K_t = k_t$ with k_t the number of connections of contiguous units associated with unit t
Σ_t = sum over all units
Σ' = sum over contiguous units

If there is no areal contiguity, the value of c will be approximately equal to one. Geary presents both a randomization and probability distribution approach to the use of the ratio for hypothesis testing.

Section 2 - Alternative Prediction Algorithms

We have selected three alternatives with which we compare TSA -- "nearest neighbor prediction", "gravity prediction", and "modified TSA". The first two methods are naive but objective in the sense that they follow some preset computation algorithm. The last alternative is a by-product of TSA but requires a certain amount of subjective judgment.

The nearest neighbor prediction approach simply uses the value of z for the geographically (linear) closest observed point to the point to be estimated. The gravity prediction model uses a weighted combination of the values of z at the

closest (measured as linear distance) four points to the predicted point and weights them by the inverse of the distance squared.

As with any regression analysis, residuals occur in TSA which can then be examined to determine if particular geographical areas are associated with positive or negative residuals. The final alternative, modified TSA, uses this information plus any subjective information available to the investigator (for example, that a particular sub-area has characteristics which differentiate it from the surrounding areas) and allows the investigator to segment the original investigation area into subares, fit new TSA surfaces to these subareas, and then predict values of z within each subarea. Of course, if the residuals exhibit little or no contiguous covariance, as measured by the ratio c defined in Section 1, the modified TSA approach to prediction is unlikely to add to the predictive power of TSA.[4]

Section 3 - Prediction Experiment

Given the basic TSA prediction method and the three alternatives outlined above, we now describe the steps followed in comparing the predictive accuracy of the various methods. The data set used is the 1970 Census of Housing for Fulton and DeKalb Counties in Georgia. The procedure to compare the predictive ability of the four methods is to select variables available at both the Census tract and Census block level and use the tract data as control points (i.e., as if these were the only data available) to predict levels of the variables for a sample of Census blocks using each of the prediction methods. The predicted values for the block level are then compared to the actual block values.

For the experiment, three different variables are used -- mean housing value, percent of persons under 18 years of age, and percent of the population which is black.

Given thse variables, the following procedure is used. From the approximately 8,000 Census blocks in the two-county area we select a random sample of 400. On a map we "eye-ball" the geographical center of each tract and the 400 blocks. Then using a cartographic digitizer we assign to each of these centers an x,y coordinate relative to a common origin.

For each Census tract the information regarding the level of the three variables listed above are extracted from the information in the 1970 Census of Housing, "Block Statistics." For those tracts for which information was not provided we use the mean of the variable in question for all tracts which are contiguous to the tract with the missing data. For the selected block the same three variables are coded; however, in this case missing data were simply excluded from the prediction error computation.

Each of the four alternative prediction methods is then used to predict values for the sampling of Census blocks. From these predictions and the observed block-level values three error measures are determined -- the sum of squared errors, the simple correlation coefficients between the predicted and actual values of the variables, and the mean percentage error where the percentage error for any observation is determined as (error/actual value).

Section 4 - Results and Conclusions

The results of the empirical tests described above are not overly encouraging to the social scientists hoping to use TSA as a prediction technique for imputing values to non-observed spatial variables. We will first consider the TSA results for the entire sample region, then compare the prediction results with those from the two naive techniques. The final portion of the section contains results on two segmentations of the orginal data space.

The first variable to be predicted is the mean housing value of a Census block. This variable would, a priori, seem to be a likely candidate for TSA under a behavioral hypothesis that persons choose to live near persons with similar housing demands and thus differences in values should vary gradually over space. (Of course, it is for this very same reason that the two naive methods may also predict quite accurately.)

The upper panel of Table 1 contains the results for the mean housing value variable for the entire two-county area. At the lower portion of that panel is the value of the contiguity coefficient, .206. The value of 1-c, .794, can be interpreted as an areal correlation coefficient, which, using the standard normal test cited above, indicates that the null hypothesis of no areal association of housing values can be rejected at less than the .001 level of significance. (The value of the standard normal deviate is 14.5. Note also that each of the contiguity coefficients reported below are highly significant.)

The total sum of squared variation about the mean for the housing value variable is shown in the lower portion of the panel with the coefficients of determination for each of the four surfaces shown in column (1). Although the first degree surface does not explain even one-quarter of the total variation, the higher order surfaces have a much higher explanatory power.

The predictive power of the surfaces are shown in columns (2)-(4) of the Table based on 302 housing value levels in the sampled Census blocks. One sees from the error measures that the second degree surface does the best predicting for this sample and this particular variable, in the sense that it produces the lowest sum of squared errors and has the highest correlation between actual and predicted values of housing values.

The two lower panels of Table 1 contain summary information on the other two variables of interest. For the age variable we find a con-

tiguity coefficient of .136 indicating an even higher areal association for this variable than for housing values. The coefficients of determination for the age variable are lower than for comparable surfaces on the housing value variable, a result not unexpected given the seemingly more random nature of this variable. For this variable, too, the second degree surface does, by far, the best job of predicting. In fact, the third degree surface results in a negative correlation between predicted and observed values of the 387 block values predicted.

The results for the racial composition variable are not encouraging either. Although, as would be expected, there is a very significant contiguity ratio, the prediction error measures do not indicate unqualified success in predicting racial composition of Census blocks on the basis of TSA using Census tract data. These results are likely affected by locational housing patterns in Atlanta where 20.1% of the Census tracts are more than 90% black while 57.5% are less than 10% black.

Before turning to the results for the alternative prediction methods, we report on a secondary finding from the TSA regression analysis of the two-county area. As other authors have noted,[5] if the original variable under investigation shows a high areal association as measured by the contiguity ratio (as each of the variables studies here do), one may investigate the contiguity ratio of the residuals from the regression to ascertain how well the regression has explained the purely areal variations of the variable.

Unfortunately, for our data, one must conclude that the spatial relationships hypothesized do not reflect well the relationships which exist. For, as is shown in Table 2, the residuals from the twelve regressions still exhibit exceedingly high (and statistically significant) contiguity effects. Thus, we might conclude that either the functional forms chosen are inadequate or that the entire model used is incorrect. For example, perhaps other explanatory variables in addition to spatial location are necessary for improved explanatory powers. However, since the purpose of this paper is simply to compare the predictive power of TSA with several naive models, we turn now to these prediction results.

Shown in columns (1)-(3) of Table 3 are the results of predicting the values of the three variables using the nearest neighbor prediction technique. As shown there, for each of the variables and for each of the error measures, this naive prediction method performed better than the TSA approach. As might be expected the correlation between predicted and actual was especially high for both the mean housing value in a Census block and the percentage of blacks living in the block. The correlation was somewhat lower for the more randomly distributed age variable.

Interestingly, as is shown in columns (4)-(6) of the table, the gravity model of prediction did not do much better than the nearest neighbor approach. In fact for housing values, prediction results, as shown by the correlation coefficient, were better using the more naive model. This is likely due to more rapid changing of mean housing values over space than changes in the racial and age compositions of the population.

For the third alternative prediction method--modified TSA-- we use two types of subjective judgments to predict block-level variables. In the first of these, purely subjective judgments about the socio-economic composition of the two-county area was used to segment the area into three subareas which we call south, central and north to refer to the approximate relative locations of the three areas. Upon this subjective segmentation, TSA surfaces were determined for each. The R^2 and correlation results are shown in Table 4 for each of the three variables and subareas. In no instance did the segmentation produce better predictions than the naive models. In some cases the R^2 statistics were higher and prediction errors lower than for the unsegmented TSA results, in other cases poorer results were obtained. This indicates that segmentation would require a variable-by-variable approach since a segmentation which might be reasonable for one variable may be entirely different for other variables.

In the second approach to modified TSA we take advantage of the capability of TSA regression programs to map the surfaces as well as residuals from the regressions.[6] It is the capability of mapping residuals which is of primary use for the modified TSA prediction technique. By studying the residuals from the original regressions it is possible to combine this objective information with a certain amount of subjective judgment in spatially segmenting the data.

We performed this operation on the racial composition variable using the results of the residual maps from the original first through fourth degree surfaces. The results of this technique are shown in Table 5. Once again the results are mixed when compared with the original surfaces; however, prediction errors are still greater for the modified TSA approach than for the naive methods.

To summarize, we have reviewed and used trend surface analysis regression techniques for predicting values of socio-economic variables and have found that, although the technique provides an objective approach to summarizing the spatial distribution of variables, it does not perform as well as alternative, simpler approaches to prediction. Included in these alternative methods have been a nearest-neighbor approach, a gravity model based on the four nearest observed points, and a modified TSA method.

We, therefore, conclude that when faced with the problem of two data sets with non-coincident boundaries, alternatives to TSA are likely to be preferred in readying the two data sets for joint analysis. The social scientist may have to give up degrees of freedom and aggregate to common boundaries or may have to use other aggregation methods.

FOOTNOTES

1/ We wish to thank Truman Hartshorn for his help and the Bureau of Business and Economic Research at Georgia State University for financial assistance.

2/ For descriptions or applications of the other techniques mentioned, see Harbaugh and Preston (1966), Tobler (1969), and Hawkes (1973).

3/ An expanded discussion of the ratio is given in Duncan (1961).

4/ The modified TSA is then simply a method for further utilization of residuals, a technique discussed by Thomas (1968). Note, it is this aspect of the problem which essentially required that the regression program employed contain mapping provisions.

5/ For example, Geary (1954) and Hawkes (1973).

6/ Note that, except for this feature, the preceding results for TSA could have been generated using any ordinary regression program. The program which we used was written by O'Leary, et al. (1964) at the University of Kansas and was adapted for use on a Univac Spectra 7 computer.

REFERENCES

Duncan, O. D., R. P. Cuzzort and B. Duncan (1961), Statistical Geography (The Free Press).

Geary, R. C. (1954), "The Contiguity Ratio and Statistical Mapping," The Incorporated Statistician, Vol. 5, pp. 114-141. Reprinted in B. J. L. Berry and D. F. Marble, Spatial Analysis: A Reader in Statistical Geography (Prentice-Hall, 1968.)

Harbaugh, J. W. and F. W. Preston (1966), "Fourier Series Analysis in Geology," Short Course and Symposium on Computer Applications in Mining and Exploration (School of Mines, U. of Arizona). Reprinted in B. J. L. Berry and D. F. Marble, Spatial Analysis: A Reader in Statistical Geography (Prentice-Hall, 1968).

Hawkes, R. K. (1973), "Spatial Patterning of Urban Population Characteristics," American Journal of Sociology, Vol. 78 (March), pp. 1216-1235.

O'Leary, M., R. H. Lippert and O. T. Spitz (1964), "Fortran IV and Map Program for Computation and Plotting of Trend Surfaces for Degrees 1 Through 6," University of Kansas.

Tobler, W. R. (1969), "Geographical Filters and Their Inverses," Geographical Analysis, Vol. 1 (July), pp. 234-253.

Thomas, E. N. (1968), "Maps of Residuals from Regression," in B. J. L. Berry and D. F. Marble, Spatial Analysis: A Reader in Statistical Geography (Prentice-Hall), pp. 326-352.

U.S. Bureau of the Census (1971), Census of Housing: 1970 Block Statistics, Final Report HC(3)-56 (Government Printing Office).

Table 1

TSA REGRESSION AND PREDICTION RESULTS

Mean Housing Value

Degree Surface	(1) R^2	(2) Sum Squared Errors	(3) Correlation	(4) Mean Proportion Error[a/]
1st	.238	.363 E11	.530	.062
2nd	.432	.308 E11	.602	.074
3rd	.485	.634 E11	.586	-.342
4th	.615	.665 E15	.462	57.018
c = .206	no. blocks for prediction = 302		sum squared variation = .160 E11	

% Less Than 18

	(1)	(2)	(3)	(4)
1st	.004	.577 E5	.032	.100 E5
2nd	.157	.508 E5	.374	.987 E4
3rd	.254	.325 E8	-.150	-.591 E5
4th	.307	.191 E9	.056	.121 E6
c = .136	no. blocks for prediction = 387		sum squared variation = .174 E5	

% Black

	(1)	(2)	(3)	(4)
1st	.101	.465 E6	.265	.206 E6
2nd	.233	.385 E6	.402	.105 E6
3rd	.256	.176 E9	.300	-.510 E7
4th	.378	.537 E11	-.257	.912 E8
c = .124	no. blocks for prediction = 357		sum squared variation = .292 E6	

[a/] Proportion Error computed as $\frac{\text{Predicted - Actual}}{\text{Actual}}$; If zero, Actual set = .0001.

Table 2
CONTIGUITY COEFFICIENTS ON RESIDUALS

Degree Surface	Housing Value		% Less Than 18		% Black	
	c	1-c	c	1-c	c	1-c
1st	.067	.933	.094	.906	.073	.927
2nd	.087	.913	.108	.892	.086	.914
3rd	.095	.905	.120	.880	.088	.912
4th	.113	.887	.128	.872	.102	.898

Table 3

PREDICTION RESULTS USING NAIVE METHODS

Variable	Nearest Neighbor (1) Sum Squared Errors	(2) Correlation	(3) Mean Proportion Error	Gravity Model (4) Sum Squared Error	(5) Correlation	(6) Mean Proportion Error
Housing Value No. used = 302	.134 E11	.845	.052 E0	.141 E11	.838	.043 E0
% $\leq$ 18 No. used = 387	.499 E5	.476	.747 E5	.473 E5	.487	.743 E4
% Black No. used = 357	.161 E6	.821	.547 E5	.146 E6	.834	.676 E5

Table 4
THREE-WAY SEGMENTATION USING TSA

Degree Surface	South R^2	South r	Central R^2	Central r	North R^2	North r
			Housing Value			
1st	.304	.042	.246	.277	.070	.544
2nd	.509	.450	.299	.302	.444	.610
3rd	.565	.083	.548	.484	.639	.580
4th	.627	.080	.589	.279	.902	-.310
Total Variation	.369 E10		.440 E10		.842 E9	
			% Less Than 18			
1st	.077	-.157	.057	.101	.594	.442
2nd	.319	.054	.250	.240	.663	.310
3rd	.414	.184	.452	.074	.804	.259
4th	.505	.170	.522	-.148	.829	.470
Total Variation	.132 E5		.114 E4		.275 E4	
			% Black			
1st	.198	.315	.242	.420	.031	.040
2nd	.459	.215	.371	.483	.046	.015
3rd	.610	-.257	.478	.392	.099	.040
4th	.620	.031	.495	-.023	.118	-.030
Total Variation	.199 E5		.205 E6		.109 E4	

Table 5
SEGMENTATION USING RESIDUALS
FROM % BLACK SURFACE

Surface	South R^2	South r	Central R^2	Central r	North R^2	North r
1st	.024	.081	.116	.319	.037	-.074
2nd	.098	.093	.354	.401	.072	-.186
3rd	.153	-.033	.405	-.287	.264	-.182
4th	.243	.098	.471	-.277	.398	.188
Total Variation	.788 E4		.159 E6		.668 E4	

ON THE EFFICIENCY OF THE SIMPLE LEAST SQUARES ESTIMATOR IN THE LINEAR MODEL UNDER $\mathcal{C}(V) \supset \mathcal{C}(X)$

Richard A. Stein, University of Arizona

1. INTRODUCTION. The linear model is often expressed as $y = X\beta + e$ where y is an $n \times 1$ vector of observations, X is an $n \times p$ matrix of known real constants, β is a $p \times 1$ vector of unknown parameters and e is an $n \times 1$ vector of errors with $E(e) = \phi$, the null vector, and with variance matrix $V = E(ee')$.

Let $SLSE(X\beta)$ denote the simple least squares estimator of $X\beta$ and $BLUE(X\beta)$ the minimum variance linear unbiased estimator of $X\beta$. This later can be expressed as $X(X'V^{-}X)^{-}X'V^{-}y$ whenever the range space of V contains the range space of X, Mitra & Rao (1968).

Durbin and Watson (1950) and Watson (1955) are among the first to lay ground work for comparing the performance of the simple least squares estimators with respect to the corresponding minimum variance linear unbiased estimators. Watson (1955) defined the efficiency of $SLSE(X\beta)$ to be $|X'X|^2/|X'VX||X'V^{-1}X|$. As this expression is the ratio of the generalized variances of $SLSE(X\beta)$ and $BLUE(X\beta)$ respectively when $X'X$ and V are invertible, it has considerable appeal. When $X'X$ or V are not invertible, certain difficulties arise.

Magness and McGuire (1962), considering $X'X$ and V invertible, establish relationships more along the lines of this paper, i.e. $\lambda_{min} \leq Var[BLUE(X\beta)] \leq Var[SLSE(X\beta)] \leq \lambda_{max}$ and $Var[SLSE(t'X\beta)] \leq (\lambda_{min}^{-1} + \lambda_{max}^{-1})(\lambda_{min} + \lambda_{max})Var[BLUE(t'X\beta)]/4$ where λ_{min} and λ_{max} are respectively the smallest and largest eigenvalues of V.

Golub (1963) extended the results of Magness and McGuire by using an inequality due to Schopf assuming invertibility of $X'X$ and V.

Rizzuto et. al. give an attainable efficiency defined as the same generalized variance ratio for the covariance structure $(1 - \rho)I + \rho J$. However, the bound of these authors appears to be attainable in the sense that for a particular covariance matrix there exists a design matrix for which their bound is attained.

Notationally, let A be any matrix. Then $\mathcal{C}(A)$ and $\mathcal{R}(A)$ are respectively the column and row spaces of A. Likewise, $\mathcal{C}^{\perp}(A)$ and $\mathcal{R}^{\perp}(A)$ are respectively the orthogonal complements of $\mathcal{C}(A)$ and $\mathcal{R}(A)$. The usual expectation and variance operators are $E(\cdot)$ and $Var(\cdot)$. Finally, A^{-} will denote any generalized inverse of A and A^{+} will be the Moore-Penrose pseudoinverse of A.

2. GENERAL APPROACH. Let us define the efficiency as

$$\min_t \frac{Var[BLUE(t'X\beta)]}{Var[SLSE(t'X\beta)]} = Eff(SLSE(X\beta)) \qquad (1)$$

Two bounds, one being exact, will be given for this definition of efficiency under the constraint $\mathcal{C}(V) \supset \mathcal{C}(X)$. The meaning of exact is in the sense that the bound is attainable for whatever specific design matrix X is chosen.

Lemma 1: $Eff(SLSE(X\beta)) =$

$$\frac{z'X'V^{+}Xz}{z'X'V^{+}X(X'X)^{-}X'VX(X'X)^{-}X'V^{+}Xz} \qquad (2)$$

where z is such that $t'X = z'X'V^{+}X$.

Such a z exists because $\mathcal{C}(V^{+}) = \mathcal{C}(V) \supset \mathcal{C}(X)$. Defining the denominator of (2) to be $f'f$, then

$$0 < 1/ \max_{z'X'V^{+}Xz=1} f'f \leq Eff(SLSE(X\beta)) \leq 1/ \min_{z'X'V^{+}Xz=1} f'f \leq 1$$

Via (2), the lower bound of the efficiency is seen to be attainable.

3. AN INEXACT LOWER BOUND.

THEOREM 1: In the model $y = X\beta + e$, $E(e) = \phi$, $Var(e) = V$ with $\mathcal{C}(V) \supset \mathcal{C}(X)$,

$$Eff(SLSE(X\beta)) \geq \frac{\text{smallest nonzero scalar } h\text{: } X'VXt = hX'Xt}{\text{largest scalar } h\text{: } X'VXt = hX'Xt}$$

This bound is, in general, not attainable for any choice of X.

It can be shown that the largest eigenvalue of any symmetric positive semidefinite matrix A must, for all choices of X, exceed the largest scalar h such that $X'AXz = hX'Xz$. Also the smallest nonzero eigenvalue of A is smaller than or equal to the smallest h satisfying this expression. Thus via the use of h, the relationship, using appropriate normalized parametric functionals,

$$\lambda_{min} \leq h_{min} \leq Var[BLUE(t'X\beta)] \leq Var[SLSE(t'X\beta)] \leq h_{max} \leq \lambda_{max}$$

provides a tightening of one of the bounds of Magness and McGuire.

4. AN EXACT LOWER BOUND.

A second approach to efficiency bounds is to

attempt to study directly $1/\text{Max } f'f$ subject to the constraint $z'X'V^{+}Xz = 1$. To this end, let the covariance matrix V be expressed in the form $V = (P|R|C)\begin{bmatrix} M & \phi & \phi \\ \phi & N & \phi \\ \phi & \phi & Q \end{bmatrix}\begin{pmatrix} P' \\ R' \\ C' \end{pmatrix}$ where M, N, Q are diagonal matrices. The columns of P form a complete set of orthonormal eigenvectors of V which lie in $\mathcal{C}(X)$. The columns of C form a complete set of orthonormal eigenvectors of V which lie in $\mathcal{C}^{\perp}(X)$. The columns of R are a complete set of the remaining orthonormal eigenvectors which are also orthogonal to $\mathcal{C}(P)$ and $\mathcal{C}(C)$.

All estimable parametric functionals can be expressed in the form $t'X\beta$. It is interesting that there exists a subspace of $\mathcal{R}(X)$ from which must come those $t'X$ for which $\text{Eff}(\text{SLSE}(X\beta))$ is attained.

THEOREM 2: In the model $y = X\beta + e$, $E(e) = \phi$, $E(ee') = V$, $\mathcal{C}(V) \supset \mathcal{C}(X)$, P, R, C as previously defined, $\text{Eff}(\text{SLSE}(X\beta))$ is attained for some parametric functional $t'X\beta$ if and only if $t'X \in \mathcal{C}(X'R)$.

The particular case where $\dim \mathcal{C}(X) - \dim \mathcal{C}(P) = 1$ has particular interest in that an attainable lower bound for $\text{Eff}(\text{SLSE}(X\beta))$ has a mathematical simplicity that seems otherwise lacking.

THEOREM 3: In the model $y = X\beta + e$, $E(e) = \phi$, $E(ee') = V$ with $\mathcal{C}(V) \supset \mathcal{C}(X)$, and $\dim \mathcal{C}(X) - \mathcal{C}(P) = 1$, let r be any vector in $\mathcal{C}(R)$ having that direction such that $\mathcal{C}(X) = \mathcal{C}(P) \oplus \mathcal{C}(r)$. Let $\{\cos \alpha_i,\ i = 1,\cdots,k\}$ be the set of directional cosines of the column vectors of R with respect to r and λ_i, the corresponding eigenvalues of V. Then $\text{Eff}(\text{SLSE}(X\beta)) = [(\sum_{i=1}^{k} \lambda_i^{-1} \cos^2\alpha_i)(\sum_{i=1}^{k} \lambda_i \cos^2\alpha_i]^{-1}$.

It should be noted that the smallest value of $\dim \mathcal{C}(R)$ is 2 whenever not all simple least squares estimators are best. This provides a setting for the following corollary.

COROLLARY: If under the conditions of Theorem 3, $\dim \mathcal{C}(R) = 2$, let $\cos\delta$ be the directional cosine of one of the two column vectors of R, and λ_{min}, λ_{max} are the two corresponding eigenvalues, then for fixed δ, $\text{Eff}(\text{SLSE}(X\beta))$ is strictly decreasing as $\lambda_{max} - \lambda_{min}$ increases. For fixed λ_{min}, λ_{max}, $\text{Eff}(\text{SLSE}(X\beta))$ is strictly decreasing as $|\delta - 45°|$ decreases.

THEOREM 4: In the model $y = X\beta + e$, $E(e) = \phi$, $E(ee') = V$, let $\mathcal{C}(V) \supset \mathcal{C}(X)$. Let P, R, C be as previously defined. Let X_2 have linearly independent column vectors such that $\mathcal{C}(X) = \mathcal{C}(P) \oplus \mathcal{C}(X_2)$ with $\dim \mathcal{C}(X_2) \geq 1$. Let ψ be any matrix where the i^{th} column vector is a set of directional cosines of the i^{th} column vector of X_2 with respect to the matrix $(P|R|C)$. Let W be any matrix satisfying $\mathcal{C}(W') = \mathcal{C}^{\perp}(\psi)$. Then $\text{Eff}(\text{SLSE}(X\beta)) = [(z'L^{+}z)(z'Lz)]^{-1}$ where L is diagonal such that $V = (P|R|C)L(P|R|C)'$ and z is any one of the solutions to $Wz = \phi$, $z'z = 1$,

$$[(z'L^{+}z)\psi'LL - 2(z'Lz)(z'L^{+}z)\psi'L + (z'Lz)\psi']z = \phi$$

COROLLARY: Under the conditions of Theorem 4, if the directional cosines of the matrix ψ are not considered, then $\text{Eff}(\text{SLSE}(X\beta)) \geq 4/(\lambda_{min}^{-1} + \lambda_{max}^{-1})(\lambda_{min} + \lambda_{max})$ where λ_{min} and λ_{max} are the smallest and largest eigenvalues of V with respect to the column vectors of R.

Here again the bounds for $\text{Eff}(\text{SLSE}(X\beta))$ of the preceding corollary are at least as close as those of Magness and McGuire since the maximum and minimum eigenvalues of the corollary are taken from a subset of all the eigenvalues of V.

5. REFERENCES

[1] Durbin, J. and Watson, G.S. (1950). Testing for serial correlation in least squares regression, I. Biometrika 37, 409-428.

[2] Golub, G.H. (1963). Comparison of the variances of the minimum variance and weighted least squares regression coefficients. Ann. Math. Stat. 34, 984-991.

[3] Magness, T.A. and McGuire, J.B. (1962). Comparison of least squares and minimum variance estimates of regression parameters. Ann. Math. Stat. 33, 462-470.

[4] Rizzuto, G.T., Boullion, T.L. and Lewis, T.O. (1973). An attainable lower bound for the efficiency of least squares estimates. 138th meeting IMS, Albuquerque.

[5] Watson, G.S. (1955). Serial correlation in regression analysis, I. Biometrika 42, 327-341.

FDU: A CASE STUDY OF THE PROCESS OF HIGHER EDUCATION[1]

Rinaldo H. Toporovsky, Fairleigh Dickinson University

A. Introduction

Social scientists have traditionally looked outside of their own community for problems requiring rational analysis and solutions. As educators, they have tried to enlighten their students on the quantitative properties of processes taking place outside of their own classrooms. However, until recently, the quantitative properties of one of their most important activities, the educational process itself, have remained outside the social scientists' purview. While refusing to heed intuition in other areas of human endeavor, social scientists have tended to rely on guesswork as a basis for decision-making within their own classrooms.

Indeed, education is a major area of public decision-making and concern. Respectable theories have been built using education as one of the main components of human capital, and thus as one of the main determinants of private income and wealth.[2] The possibility of affecting the income and wealth distribution of the population by means of rational educational policies thus becomes clear and desirable. Nevertheless, educational levels cannot be controlled unless the learning process is clearly understood.

In the last few years, economists have produced a large and evergrowing, collection of "production-function" studies. These studies have tried to establish how resources are allegedly used in the attainment of alternative levels of production. Functions having a priori desirable properties, such as the C-D and CES production functions have been called upon to link inputs and outputs in specified ways. However, it is well known that the power of these hypotheses is small; alternatively highly probable explanations usually offer conflicting views of the empirical findings. In the field which concerns us now, input-output relations, having uncertain empirical properties, have multiplied in number -- leaving the educational researchers with the same unanswered questions that baffled their methodological predecessors.

Other social scientists, having developed no vested interests in the empirical viability of partial technological relations, have tried to explain the process of education in terms of socio-economic determinants. Unfortunately, the latter type of studies have tended to disregard the a priori properties of their postulations -- increasing the likelihood of confounding.

In this paper, I will develop a theory of educational decision-making, and summarize its empirical implications upon a specific sample of data on higher educational variables, drawn from Fairleigh Dickinson University's student population.[3] The theoretical approach will focus upon the implications of constrained decision-making. The observed differences in student achievements, or educational outcomes, will not only be connected with technological and endowment differences, as it is true of the traditional "production-function" studies, but also with aspirations, or goal variations. However, the interactions between goal and technological relations will be established iteratively, in order to determine the separate impact of each of the two sets of variables respectively. The empirical approach will be Bayesian in nature, in the sense that the data will be used to infer the theoretical classifications. In other words, rather than "explain" achievement levels, I will use the observed levels of achievement in order to determine the likelihood of the postulated theory.

B. An Exact Decision-Making Theory

1. The empirical manifestations of the relations existing between educational inputs and outputs are not independent from the goals and objectives of the individuals and institutions involved. It is only commonsensical to link student achievement to own evaluations of education -- or to the educational policies of the colleges in question. Thus, unless the crucial role played by goals and objectives is clearly recognized at the outset, any predictive statement of student behavior, failing to take it explicitly into account, will be subject to an uncertain degree of bias.

2. In a restricted sense, given our ignorance of the student's spectrum of goals, resources and technological constraints, we could linearly approximate the decision-making problem of the student as follows:

$$\max U = u'x \text{ s.t. } (I\ Z : II x \underset{\sim}{\leq} x_{\sim}^{\circ}) = b, \text{ where}$$

u'= vector of individual, ordinal prefer-

ence weights with respect to the various x, U= scalar representing level of goal attainment, x= vector of outputs -- educational, recreational, work, or goals of the individual student, Z= matrix of educational, recreational, work technology coefficients, I= identity matrix, b= vector of given resource constraints, s= vector of slack variables associated with the vector of endowments b.

Consider Z_B as the optimal basis for the problem at hand, such that x*, c* are optimal if and only if $x^*_j=0$ for $z'_j c^* > u_j$, and $c^*_i=0$ for $z_i x^* < b_i$.[4] These conditions imply that students will neither choose to produce any output x^*_j for which its ordinal returns per unit -- the price of x^*_j, u_j -- are smaller than its imputed costs, $Z_j c^*$, nor consider any resource which remains unused valuable.

At the optimum, which is an extremum, or a set of extrema, $Z_B x^* = b$ or $Z_B^{-1}(b) = x^*$. Thus, for any output $x^* = (Z_B^{-1})x\ b$, and $x^* > 0$ if and only if $\Sigma_j (z_{1_j} c_j) - \mu_1 < 0$ -- where the subscript $_{1j}$ indicates the matrix element in row 1 and column j. These necessary and sufficient "marginal" conditions could also be restated as follows:

$(Z_B)j = (\mu_i / c_j)$, j= 1, 111, n.

If we let $\overset{n}{x}_1^*$ stand for a particular educational output, and b for a column vector whose elements are the levels of a chosen set of educational resources, we may state the strong hypothesis that educational outputs are a function of educational inputs. This is a strong hypothesis because we are assuming both a specific linear form for our approximation and a given evaluation of educational outcomes. Moreover, assuming that students decisions approach optimality, we will conclude that those resources which do not appear to be used in the achievement of a particular educational goal have a negative yield -- returns on their use are less than their costs.

Take, for instance, cumulative grade averages. Students having a given level of intelligence, faculty assistance, library use, etc. will achieve, according to our postulation, the same grade average if their evaluation of different grade averages is consistent as well. It is worth emphasizing that two students with the same resources may achieve different grade levels because their goals are dissimilar.

C. Statistical Methodology: Parameter Estimation

1. Our methodology will be perfectly general, and thus applicable to any perceived educational output. Since our preliminary empirical tests concentrate on cumulative grade averages, we will phrase our statistical considerations in terms of those variables -- without loss of generality and for ease of understanding. According to our hypothesis, student cumulative grade averages, as an educational output, may be expressed as a linear combination of student resources, provided the subjective, relative evaluation of the importance of grades is maintained constant within each grade group. Empirically, this hypothesis implies that we should expect students in different cumulative grade average classes to exhibit, respectively, characteristic patterns of resource endowments and use. For each grade level x_1 -- F, D, C, B, A, -- we will compute a vector $(Z^{-1})_{1.}$ such that the resulting sample scalar $(Z^{-1})_{1.} b_{.a} = x_{1a}$ for each of the students, a= 1, 2, ..., n_1, in that grade class will minimize the probability of classifying that student, a, in a grade level other than that in which he is observed. $b_{.a}$ is an (mxl) column vector of m resource levels for student a. In other words, given several groups of students exhibiting grade averages ranging from F to A, we want to determine a set of constants for each group such that the corresponding linear combinations of student resources will yield for each group respectively a range of scalars which minimize the probability of misclassification.

2. Consider the following matrix of observations: $B = /b_{ij}/$, with typical element b_{ij}, where i= resource variable under considerations, i= 1, 2, ..., r, and j= ordered student number, j= 1, 2, ..., $\sum_{l=1}^{p} n_l$. Order the columns of B in such a way that the first n_1 correspond to the lowest grade students, n_2 to the next to the lowest grade, etc. up to the last n_p observations -- corresponding to the highest cumulative average grade students. Define S=

$$(BB' - \sum_{l=1}^{p} (n_l \bar{b}\bar{b}')) / ((\sum_{l=1}^{p} n_l) - p),$$

where $\bar{b}(\bar{b}_1 \bar{b}_2 \bar{b}_3 \ldots \bar{b}_m)'$, and $\bar{b}_i = (b_{i.} xI) / \sum_{l=1}^{p} n_l$ -- I is an $(\sum_{l=1}^{p} n_l)$xl column vector of ones -- as the unbiased or pooled covariance matrix for the whole sample.

Then, for each grade group, the maximum likelihood estimates of $(Z^{-1})_{1.}$ will be as follows:

$(Z^{-1})_{1.}=((-\frac{1}{2}\ \bar{b}_{.1}'S^{-1}\bar{b}_{.1})\ (\bar{b}_{.1}'S^{-1}))$,[5]

where $\bar{b}.=(b_1'\ I/\ n_1)\ (b_2'I/n_1)\ \ldots$

$(b_m/I/n_1)$ and I is a $(\sum_{1=1}^{p} n_1 x1)$ vector,

with ones corresponding to students who got grades x_1, and with zeroes elsewhere. It is worth reminding the reader that $(Z^{-1}_{1.})\ -\ (Z^{-1})_{k.}=D_{1k}$ for $k \neq 1$ represents the discriminant function coefficients between groups 1 and k. In this particular case, we have a total of $(p^2-p)/2$ discriminant functions.

3. Some observations will appear to be inconsistent, in the sense that, according to resources, students will be classified in grade groups different from those in which they were actually observed. We may infer, from our previous analysis, that such inconsistencies are due to goals differences. We strongly hypothesize that, for instance, all A students observed classified as B, C, D, or F, according to their resource endowments, attained the highest grade as a consequence of their different evaluation of grade outputs. In order to identify such differences, we will associate goal levels with a relevant set of socio-psychological background variables. In symbols, we will consider the following linear approximation:

$u^a_{11}*= K.g^a$ where $u^a_{11}*=$ value of cumulative grade average to student a, who attained a level x_1* but appeared, in terms of resources, to belong to the group x_1, 1=A, B, C, D, F, K=(1xm) vector of marginal, constant goal contributions, m= total number of socio-psychological background variables, p= total number of grade groups, g^a= (mx1) vector of socio-psychological background variable levels corresponding to student a.

In passing, it may be noted that an ordinal notion of the marginal costs of the resources employed by students in each grade classification may be gained from the following relation:

$(Z^{-1})'_1*u'_{11}*=c_{.1}*$, which defines, given the programming equilibrium conditions, the implied cost of such resource endowment, at the margin, for a student attaining a grade level 1*, but belonging to the resource group 1. In general, since $K.g= u_{11*}$, we may state that $c_{.1}*= ((Z'^{-1})_{1*}\ K.\bar{g})$, for each 1, 1*, K triplet applicable.

4. In sum, the statistical methodology which will be applied, in order to identify endowment and goal influences at each cumulative grade level may be summarized as follows:

(a) Classify students according to cumulative grade levels.

(b) Estimate a set of "marginal product" coefficients for each group, in order to explain grade differences in terms of resource endowments and utilization.

(c) Estimate a set of "marginal goal contribution" coefficients for each group of similarly misclassified students within each originally observed grade, in order to explain switches in terms of socio-psychological background variables.

5. In the first iteration, once the inputs considered of importance have been isolated for each grade group, it becomes important to determine which variables appear to be instrumental in promoting students to cumulative grade averages other than that which they have attained. Given five grade levels, techniques of producing grades, there are ten possible distinct movements we may examine. In other words, we may ask the following types of questions: if, according to observed results, a student attained a B average, and he wants to promote his grade to the A level, should he increase hours of library study, and/or cut down part time work activities, and/or seek enlarged faculty assistance, etc.? Similarly, in the second state or iteration, we want to determine those socio-psychological background variables which appear to induce grade-attainment inconsistencies.

Statistically, those marginal "expansion" coefficients can be defined as: $S^{-1}(\bar{b}_{.1}-\bar{b}_{.k})=((Z^{-1})_{.1}-(Z^{-1}{}_{.k}))$ for all groups, 1≠k, and upon pairwise comparison.[6]

D. Statistical Methodology: Sample Distribution

1. Total Distance

Whether the optimal classification is statistically significant, or just a "figment of sampling variation" may be determined in terms of the distance between the mean vectors of the optimal groupings and that of the total sample. P. C. Mahalanobis'[7] generalized distance measure may be applied, and its significance determined, for samples of the size we will consider. For example, for the first iteration, on educational re-

sources,

$M=\sum_{i=A}^{F}(\bar{b}_1-\bar{\bar{b}})' S^{-1}(\bar{b}_1-\bar{\bar{b}})$ where M= generalized Mahalanobis distance, $\bar{b}$= vector of resource means for grade level 1, $\bar{\bar{b}}=(\bar{b}'i/p)$, p= total number of resource groups -- as the reader may recall from the previous discussion -- i= a vector of ones (1xr), and r= number of resource constraints at the optimum. M has a χ^2 distribution with px(r-1) degrees of freedom.

2. Intergroup Distances

Testing whether the groups' mean vectors of resources are significantly different from the sample's total mean resource vector or not gives us a general idea of the statistical validity of our hypothesis. However, we may want to test the statistical significance of the (px(p-1)/2) distances among the mean resource vectors, upon pairwise comparison, of the p groups. In this regard, we may recall that

$$T^{*2}=(\bar{b}_1-\bar{b}_k)'S^{-1}(\bar{b}_1-\bar{b}_k).\ ((n_1+n_k)-1)_{1\neq k},$$

where T^{*2}= Hotelling's T^2, which transforms, upon multiplication by $(n_1+n_k-r)/r(n_1+n_k-1)$ into a central F^8 variable with r and (n_1+n_k-r) degrees of freedom, and $\bar{b}_1$ and $\bar{b}_k$ represent groups' 1 and k mean resource vectors respectively.

Thus, $F_{1k}=2([(Z^{-1})_{.1}-(Z^{-1})_{.k}]'\bar{b}_1).$ $(n_1+n_k-r)/r$ and the (px(p-1)/2) hypothesis may be tested at any desired level of significance respectively upon comparison of F_{1k} and F^{α}, where F^{α} represents a critical calue for a significance region of size α.

3. Posterior Probabilities

The probability that a given student will be considered well-classified or not may be determined in a "Bayesian" fashion. For instance, consider the first discrimination iteration, in terms of resource endowments. In particular, for a student a, who achieved an x_2 average, five values, x_1= A, B, ... F, $(Z^{-1})_{1.} \times b_{.a}= f^a_{x1}$ may be computed, and the maximum $f^a_{x1}= f^M$ determined. Then, we may establish a priori, that $P(x_2/x_1).P(x_1)= g(f^M - f^a_{x1})$, choosing a function g which will achieve a maximum at g(0), and such that $0\leq g<1$. The exponential $\exp(-(f^M - f^a_{x_1}))$, for instance, fulfills our requirements. Thus, the posterior $P(x_1/x_2)$, for 1= 1, ..., 5, may then be computed using Bayes' theorem.

Alternatively, since $(f^a_{x1}-f^a_{xk})= N((\bar{f}_{x1}-\bar{f}_{xk}),\ 2(\bar{f}_{x1}-\bar{f}_{xk}))$, $1\neq k$, $\bar{f}_{xk}=((Z^{-1})k).\ x\bar{b}_{.k}$, $\bar{f}_{x1}=(Z^{-1}{}_{1.}).\ x\ \bar{b}_{.k}$ if x_k is true, and $(f^a_{xk} - f^a_{x1})= N(-\bar{f}_{xk}+\bar{f}_{x1}),\ 2(\bar{f}_{xk} - \bar{f}_{x1}))$, $k\neq Z$, when x_1 is true, pairwise critical regions may be established for each student vector.

4. Dependence and Prediction

Whether the postulated hypothesis: (1) patterns of educational resource use and outcomes are functionally related, and (2) inconsistencies of resource use and outcomes are due to educational goal differences, are empirically corroborated or not may be tested in alternative ways.

Contingency tables indicating resource classification frequencies for each alternative outcome, average cum grades, and goal classification frequencies for each alternative resource classification within a particular outcome group may be constructed. Then, Pearson's χ^2 approximation may be used to test the independence of (1) specific, or optimal patterns of resource use from their corresponding outcomes, and of (2) specific, or optimal patterns of social background variables from that of alternative resource use classification within a particular educational outcome group.

If resource use patterns can be used to predict educational outcomes, a significantly positive correlation between typical resource patterns and observed outcome classifications will exist. Analogously, given any specific educational outcome level, say average cum grades, resource use levels inconsistent with those typical for the average cum in question will have to be explained by systematic differences in social background or goal variable patterns. Thus, similarly "inconsistent" resource vectors will be positively correlated with social-background variable patterns. The Student-t distribution may be used to determine the significance of the sample correlation coefficients between educational outcomes and resource-use classifications, as well as between resource-use and social background, or goal classification within a particular educational outcome group.

The relative frequency of accurate prediction of attained levels of educational achievement may be used as an indication of the explanatory ability of

the postulated hypotheses. I call this ratio the consistency coefficient $E = \Sigma_{ij}(f_{ii}/\Sigma_{ij}f_{ij})$, where f_{ij} = number of students attaining an original classification level i, and whose discrimination, or optimal new classification level is j.

E. Summary of Empirical Findings

Our analysis of the Fairleigh Dickinson University questionnaire evidence leads to the following empirical statements:

(1) Statistically significant, distinctive educational resource use vectors characterize student average-cum grade achievements.

(2) Statistically significant, distinctive social background or goal vectors characterize specific differences between student educational resource use patterns and their corresponding average-cum grade achievements.

(3) The optimal patterns of educational resource use and average-cum grade achievements are dependent and positively correlated. Both, dependence and correlation are statistically significant.

(4) The optimal patterns of educational background and resourse use variables, within particular average-cum groups of students, are dependent and positively correlated. Both, dependence and correlation are statistically significant.

(5) The explanatory ability with respect to student average-cum grades of educational resource use and social back-ground patterns ranges from a third to more than one half of all observations.

(6) At different average-cum grade levels, definite patterns of resource use and goal variable substitution are observable. Specific grade levels may be attained in a number of resource use and social background or goal variable level combinations.

(7) The analytical methodology developed appears to be robust, in the sense that reversing the order of iteration, or changing the number of components in the explanatory vectors will not substantially alter the empirical conclusions.

Some qualifications to our findings are certainly in order. The general validity of our quantitative parameter estimates may be impaired by the nature of the sample considered. The biases introduced by probable self-selectivity or respondents, subjective content of graphic forced-choice questions, imperfect proxies, restrictiveness of sample, etc., though unknown, should not be ignored. Clearly, further testing of the theory on more comprehensive bodies of data will serve the purpose of generalizing our preliminary research. However, the empirical implications of our conclusions are strong as well as eminently practical: the process of education and educational achievement can be quantitatively assessed and the resulting parameter estimates used to devise efficient plans for university, student, and faculty resource utilization.

Footnotes

1. This paper was written under an F.D.U. Faculty Grant. I would like to specially acknowledge the advice of Dr. A. Jaffe, the data processing help of Mrs. Louise Yanoff, and the expert typing of Mrs. Marilyn Meyers.

2. Among the very many references, two are worth mentioning; Friedman, M., "Choice, Chance, and the Personal Distribution of Income," The Journal of Political Economy, vol. vli, #4, 1953, pp. 277-9-, and Becker, G.S., Human Capital, New York, 1964.

3. Lack of space prevents a full reporting of my preliminary findings. However, I will be happy to supply the interested reader with detailed statistical results.

4. C is a vector of imputed costs of resources; this is the so-called equilibrium theorem of linear programming. See, for instance, Lancaster, K. J., Mathematical Economics, New York, 1968, pp. 33-34.

5. See, for instance, Anderson, T. W.,

An Introduction to Multivariate Statistical Analysis, Chapter 6, New York, 1958.

6. For a development of this discrimination concept see, for instance, Hoel, P. G., Introduction to Mathematical Statistics, Third Edition, New York, 1964, pp. 179ff.

7. See Mahalanobis, P. C., "On the Generalized Distance in Statistics," Proceedings of the National Institute of Sciences, Vol. xii, Calcutta, India, 1936, pp. 49-55.

8. For a derivation of the distribution of Hotelling's T^2, see Anderson, T. W., Op.cit.

THE EFFECT OF STRATIFICATION WITH DIFFERENTIAL SAMPLING RATES ON ATTRIBUTES OF SUBSETS OF THE POPULATION

Joseph Waksberg, Westat, Inc.

This paper discusses the strategy to follow in stratification when one is interested in the attributes of subsets of the population, and the subsets cannot be isolated from the general population, in advance of the sampling. It pulls together the theory relating to this problem, provides data for several practical examples, and discusses some of the implications.

An example can best illustrate how the sampling issues arise. Suppose one is interested in attributes of a specific subgroup of the population, e.g., of negroes, low-income families, preschool age children, etc., but the only frames available for sampling comprise the total population and the subsets cannot be determined except as part of the interviewing procedure. A common strategy is to use geographic stratification, classifying such areas as tracts or Census EDs by the proportion of their populations in the specified subgroups. Census data may be used for stratification or more current local knowledge, if that is available.

More specifically, this paper explores the reduction in sampling variance that is possible when: (a) the population is divided into two strata in such a way that one stratum has a considerably higher proportion of the subset of interest than the other stratum, and (b) a higher sampling rate is used in the stratum with the greater concentration. Further, the paper is restricted to situations in which the following conditions apply:

(1) The stratum with the higher concentration contains less than half of the total population.

(2) Simple random sampling is used with a rate small enough so that the finite population correction factor is trivial.

(3) Most of the discussion relates to cases where the population variances in the subset are the same in both strata.

The first condition is fairly minor since, when it does not apply, only trivial gains in the variance are usually possible. In most cases, the second condition should also lead to only a minor loss of generality. The third condition is more troublesome. Situations exist in which the variances can be expected to be different. A re-examination of the major results would have to be made under such conditions, since it is difficult to state general principles when this occurs.

I. Notation and Fundamental Relations

Assume the population is divided into two strata.

N_1 or N_2 = population of stratum 1 or stratum 2.

$N_2 = vN_1$, where $v \geq 1$.

t_1 or t_2 = proportion of stratum 1 or 2 in specified subgroup.

$t_1 = ut_2$, where $u \geq 1$.

σ^2 = population variance of a statistic within the subgroup, identical in the two strata.

r_1 or r_2 = sampling rate in stratum 1 or stratum 2.

$r_1 = kr_2$, where $k \geq 1$.

Compare two sampling plans:

A: Uniform sampling rate in the two strata, rate = r.

B: Use of r_1 in stratum 1 and r_2 in stratum 2, with

$$r(N_1 + N_2) = r_1N_1 + r_2N_2$$

so that the total sample sizes are identical in both plans.

Using the usual approximations to the variance, and assuming the finite correction factors are trivial, the variance of sample means can be expressed as:

$$\sigma^2 \text{ (plan A)} \equiv \sigma_A^2 \doteq \frac{\sigma^2}{r(t_1N_1 + t_2N_2)} = \frac{\sigma^2}{rt_2N_1(u + v)}$$

$$\sigma^2 \text{ (plan B)} \equiv \sigma_B^2 \doteq \frac{\sigma^2}{r_2(t_1N_1 + t_2N_2)^2}\left[\frac{t_1N_1}{k} + t_2N_2\right]$$

$$= \frac{\sigma^2(\frac{u}{k} + v)}{r_2t_2N_1(u + v)^2}$$

$$\sigma_B^2/\sigma_A^2 \doteq \frac{k + v}{k(1 + v)} \cdot \frac{u + kv}{u + v}$$

II. Condition for $\sigma_B^2 < \sigma_A^2$

$\sigma_B^2/\sigma_B^2 < 1$ when $k < u$ -- that is, oversampling in stratum 1 will reduce the variance provided that the extent of oversampling is less than u, the ratio of the concentration of the subset in stratum 1 to stratum 2.

III. Minimum Value of σ_B^2/σ_A^2 for a Given Set of Values of u and v

For a given set of values of u and v, the optimum value of $k = \sqrt{u}$.

For this value of k, σ_B^2/σ_A^2 is equal to

$$\frac{(\sqrt{u} + v)^2}{(1 + v)(u + v)} .$$

Table 2 shows the size of this ratio for selected values of u and v.

IV. Minimum Value of σ_B^2/σ_A^2 for a Fixed Value of u

For a given value of u, the minimum value of σ_B^2/σ_A^2 occurs when $k = v = \sqrt{u}$.

When this occurs, the ratio σ_B^2/σ_A^2 is

$$\frac{4\sqrt{u}}{(1 + \sqrt{u})^2} .$$

Table 1 shows this minimum for a range of values of u.

In practical situations, it is not possible to manipulate the value of v. Once u is determined, this automatically fixes v. However, it is useful to be able to examine the minimum variance that can occur under the best possible situation.

V. Value of σ_B^2/σ_A^2 When the Population Variances Are Not Identical in the Two Strata

If the variances in the two strata are not identical, let

σ_1^2 or σ_2^2 = population variance in stratum 1 or stratum 2.

$$\sigma_1^2 = w\sigma_2^2 .$$

In this case:

(1) $$\sigma_B^2/\sigma_A^2 = \frac{(k + v)(uw + kv)}{k(1 + v)(uw + v)} .$$

(2) σ_B^2/σ_A^2 will be less than 1 when $k < uw$.

(3) The minimum value of σ_B^2/σ_A^2 occurs when $k = \sqrt{uw}$. When this occurs, the value of σ_B^2/σ_A^2 is

$$\frac{(\sqrt{uw} + v)^2}{(1 + v)(uw + v)} .$$

VI. Discussion

1. The reductions in variance will be fairly small unless the concentration of the subset of the population in the stratum to be oversampled is considerably greater than in the rest of the universe. For example, if the concentration in one stratum is twice as great as the other and the variances are the same within the two strata, at best a four percent reduction in the variance can be attained. If the concentration is four times as great, the maximum reduction is 11 percent, and then only if the ratio of the populations in the two strata turns out to be exactly 2 to 1. More likely, the gains will be in the five to ten percent range. When the concentrations get to be of the order of 10 to 1, then sizable reductions occur.

2. On the basis of the preceding comments, it is possible to assess the value of geographic stratification for many types of statistics. For example, it is unlikely that oversampling for such populations as school age children, women of child bearing age, or older persons would have any important payoff. A cursory examination of tract statistics does not reveal any important differences in age distributions among tracts, except in a trivially few tracts. The best one might expect from a stratification of tracts is probably a factor of two or three in the concentrations. Census ED's would be somewhat better, but not strikingly so.

On the other hand, oversampling to produce Negro statistics could produce useful reductions in the variances, and the same is true of low income households although to a lesser extent. A two-way stratification of high and low Negro concentrations by the Bureau of the Census, using 1960 ED's as the units of stratification and 1960 data to classify the ED's shows that the maximum reductions in variance could be in the range 30-50 percent, depending on how current were the data used for stratification. For statistics on low-income households, poverty areas defined on the basis of 1960 data would have produced a 15 percent reduction in variance, about ten years later. Presumably, if smaller areas such as tracts or ED's had been used, the reduction would have been greater, possible of the order of 20-25 percent.

3. It is somewhat deceptive to use Census data some years after the Census and assume the same efficiency applies. For Negro statistics, for example, the

values of u typically dropped by about half between 1960 and 1967, for ED's classified on the basis of 1960 characteristics, resulting in only about two-thirds of the reduction in variance that might have been expected.

4. This deterioration over time in the effectiveness of stratification for many social and economic characteristics will frequently occur even when one uses what would appear to be better modes of stratification than geographic areas. For example, assume that statistics on low-income families is desired, and it is possible to stratify individual families on the basis of the previous year's income. CPS data on the proportion of families that changed their poverty status between 1964 and 1965, indicates that 31 percent of the 1964 poor were nonpoor in 1965, and eight percent of the nonpoor became poor. The values of u and v are about five and nine. Thus the reduction in variance that would occur with the optimum k is only 24 percent. This is not much better than would result from geographic stratification.

5. Classifying the population into two strata will for most cases provide most of the gains that stratification can produce. It would take a very unusual distribution of the population, for additional strata to reduce the variance much further. This can be seen most easily by starting with a two-way stratification and examining the effect of splitting each stratum further. It is clear from comments made earlier that important gains will occur only if there are sizable differences in the concentrations of the subsets in the two substrata formed from each of the original strata. If the original stratification was reasonably effective, it would be highly unusual for substratifications to produce additional differences in concentration of 5 or 10 to 1, these being differences that are required for important reductions in variance.

6. It should be noted that all of the discussion is related to attributes of subsets of the population. If one is interested in estimates of the sizes of the subsets, the same reductions do not apply. In fact, under some circumstances, the optimum sampling rates for the attributes will result in an increase in variance over a uniform sampling rate.

7. Tables 3 and 4 indicate the values of u and v that can be expected for kinds of items for which geographic stratification is most effective -- characteristics of the Negro and low-income population. Table 5 shows the deterioration over time in effectiveness when the population is stratified into poor and nonpoor families.

APPENDIX

I. Fundamental Relations

Since

$$N_2 = vN_1$$

$$t_1 = ut_2$$

$$r_1 = kr_2$$

$$\sigma_1^2 = w\sigma_2^2$$

and

$$\sigma_B^2 = \left(\frac{t_1N_1}{t_1N_1 + t_2N_2}\right)^2 \frac{\sigma_1^2}{r_1t_1N_1} + \left(\frac{t_2N_2}{t_1N_1 + t_2N_2}\right)^2 \frac{\sigma_2^2}{r_2t_2N_2},$$

replacing N_2, t_1, etc. by their values above

$$\sigma_B^2 = \frac{1}{(t_1N_1 + t_2N_2)^2}\left(\frac{\sigma_2^2t_2N_1}{r_2}\right)\left(\frac{wu}{k} + v\right).$$

For plan A, k = 1, and r_2 is replaced by r.

Since

$$r_1N_1 + r_2N_2 = rN$$

and

$$N_1 + N_2 = N\ ,$$

replacing N_2 by vN_1 and r_1 by kr_2, it follows that

$$r = r_2\,\frac{k + v}{1 + v}\ ,$$

which leads to

$$\sigma_A^2 = \frac{1}{(t_1N_1 + t_2N_2)^2}\,\frac{(\sigma_2^2t_2N_1)(1 + v)}{r_2(k + v)}\,(wu + v)$$

and

$$\sigma_B^2 \Big/ \sigma_A^2 = \frac{(wu + kv)(k + v)}{k(1 + v)(wu + v)}\ . \qquad (1)$$

When the variances in the two strata are equal w = 1, in this case

$$\sigma_B^2 \Big/ \sigma_A^2 = \frac{(u + kv)(k + v)}{k(u + v)(1 + v)}\ . \qquad (2)$$

II. Optimum Value of k

Differentiating equation (1) with respect to k and equating to zero, results in

$$k = \sqrt{uw} \ . \tag{3}$$

With this value of k, equation (1) becomes

$$\text{Minimum } \sigma_B^2/\sigma_A^2 = \frac{(\sqrt{uw} + v)^2}{(1 + v)(uw + v)} \ . \tag{4}$$

When the variances in the two strata are equal and w = 1, equations (3) and (4) become

$$k = \sqrt{u} \tag{5}$$

$$\text{Minimum } \sigma_B^2/\sigma_A^2 = \frac{(\sqrt{u} + v)^2}{(1 + v)(u + v)} \ . \tag{6}$$

Table 1. Minimum Value of σ_B^2/σ_B^2 for Specified Values of u

u	Optimum value of k & v	Minimum of σ_B^2/σ_A^2
1	1	1.00
2	1.4	.97
4	2	.89
9	3	.75
16	4	.64
25	5	.55
49	7	.44

Table 2. Minimum Value of σ_B^2/σ_A^2 for Specified Values of u and v

u	Optimum value of k	Minimum value of σ_B^2/σ_A^2 when v = :									
		1	2	4	6	8	12	16	24	30	50
1	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	$\sqrt{2}$	.96	.96	.97	.98	.98	.99	.99	.99	.99	1.00
4	2	.90	.89	.90	.91	.93	.94	.95	.97	.97	.98
9	3	.80	.76	.75	.77	.80	.82	.85	.88	.90	.93
16	4	.74	.67	.64	.65	.67	.70	.74	.78	.81	.87
25	5	.69	.60	.56	.56	.57	.60	.63	.69	.72	.79
49	7	.64	.53	.46	.44	.44	.46	.48	.53	.56	.64
100	10	.60	.47	.37	.35	.34	.34	.34	.37	.40	.47

Table 3. Use of 1960 Census Data for Stratification of E.D.'s for the Nonwhite Population; Effectiveness at Time of Census and Seven Years Later

Geographic area	Percent nonwhite		Enrichment factor U	Percent of total pop.		Ratio of residual stratum pop. to nonwhites stratum pop. V	Percentage reduction in variance with optimum k
	In nonwhite stratum	In residual stratum		In nonwhite stratum	In residual stratum		
Data for 1960							
SMSA's 1,000,000+							
N.E.	69.7	2.6	27	12.1	87.9	7	45
N.C.	85.6	1.5	57	13.7	86.3	6	62
S	77.3	1.7	45	25.5	74.5	3	49
W	47.9	1.4	34	16.4	83.6	5	48
SMSA's 250,000-1,000,000	85.3	3.0	28	9.9	90.1	9	44
SMSA's <250,000	63.9	1.6	40	13.7	86.3	6	51
Balance	56.0	6.6	8	8.2	91.8	11	21
Data for March 1967							
SMSA's 1,000,000+							
N.E.	81.8	5.2	16	8.7	91.3	10	31
N.C.	90.5	6.7	14	10.5	89.5	9	28
S	82.3	6.9	12	19.1	80.9	4	30
W	68.8	4.2	16	11.8	88.2	7	34
SMSA's 250,000-1,000,000	92.6	6.4	14	7.0	93.0	13	25
SMSA's <250,000	55.6	2.5	22	10.1	89.9	9	39
Balance	52.0	7.5	7	7.4	92.6	13	13

NOTE: Data based on stratification performed by the U.S. Bureau of the Census for a special survey performed for the O.E.O. Data for the top half of the table are from the 1960 Census; data for the lower half are from the special survey (SEO).

Table 4. Effectiveness of Using Poverty Areas as Strata for Families in Poverty /1

Percent in Poverty	
In Poverty Areas	14.8 percent
In Non-Poverty Areas	2.6 percent
Enrichment Factor	u = 6
Percent of Total Pop.	
In Poverty Areas	14.4 percent
In Non-Poverty Areas	85.6 percent
Ratio of Total Pop. in Non-Poverty to Total Pop. in Poverty Areas	v = 6
Reduction in variance with optimum k	15 percent

1 Poverty areas are defined on basis of 1960 Census data, and restricted to SMSA's of over 250,000 population. The population distributions shown are for 1968.

Table 5. Poverty Status in 1964 and 1965 for Matched Families /1

Classification in 1965	Classification in 1964		
	Total	Poor	Nonpoor
Number of Cases (in 000)			
Total	43,845[2]	7,621	36,224
Poor	7,968	5,246	2,722
Nonpoor	35,877	2,375	33,502
Percent Distribution			
Total	100	100	100
Poor	18	69	8
Nonpoor	82	31	92
v = 4.8			
u = 8.5			
Maximum reduction in variance = 24 percent			

1 Source: special tabulations of March 1964 and 1965 CPS records.

2 The number of matched families is less than the total number of families because of births, deaths, migration, and changes of family composition between 1964 and 1965.

MULTIVARIATE ANALYSIS THROUGH A SYMMETRIC PARTIAL GAMMA

J. Sherwood Williams and B. Krishna Singh
Survey Research Center, Virginia Commonwealth University

The symmetrical measure of association proposed by Goodman and Kruskal (4) called gamma (γ) has proven to be a useful and increasingly popular measure of ordinal association in situations where x and y are ordered polytomies. This measure indicates how much more probable like orders are than unlike orders in two classifications when two cases are selected at random (4). Costner (2) has proposed an even clearer interpretation for gamma based upon the "proportional-reduction-in-error" (PRE) criterion.

Often researchers wish to examine the association between two polytomies while controlling for other polytomies. Goodman and Kruskal (4) suggested that measures of partial association for gamma might be developed for both asymmetrical and symmetrical situations. Davis (3), following Goodman and Kruskal, has developed two asymmetric measures of partial gamma. The first measure is based upon a weighted average of the conditional gamma coefficients in the different strata. The formula suggested by Davis may be written as:

$$\gamma_{xy.z} = \frac{\sum_{i=1}^{k} \gamma_{xy.z} \ (\Pi_s + \Pi_d)}{\sum_{i=1}^{k} (\Pi_s + \Pi_d)} \qquad (1)$$

where x, y, and z are ordered polytomies; Π_s denotes like orders and Π_d denotes unlike orders.

The other measures suggested by Davis is "based directly on probabilities of error." The second formula is:

$$\gamma_{xy.z} = \frac{\Pi_s xy.z - \Pi_d xy.z}{\Pi_s xy.z + \Pi_d xy.z} \qquad (2)$$

Formula 2 indicates how much more probable it is to get like orders in measures x and y when pairs of individuals differing on x and y and tied on z but unselected on any other measure are chosen at random from the population (3).

It is relatively simple to compute sample frequencies corresponding to $\Pi_s xy.z$ and $\Pi_d xy.z$ when the variables are dichotomies or trichotomies. However, difficulties arise in situations where the number of ordered categories in the polytomies increase beyond three or when the sample is not large. The risk of having zero cells, which would falsely raise the value of the partial coefficient, is increased as the number of categories in each variable increases. For example, if we have three ordered trichotomies, we have 27 cells in our cross-classification. If our three ordered polytomies have five categories each, the number of cells is increased to 125. Needless to say, the problems faced in terms of computation and interpretation increase rapidly as the number of cells in the cross-classification increase, particularly if we encounter zero cells. Further, the computation becomes even more cumbersome when it is necessary to go beyond the first order partials. Also, if the causal linkages and time order of the variables are unclear, it is not logical to compute an asymmetrical partial association.

The intent of this paper, then, is to arrive at a symmetrical partial coefficient for Goodman and Kruskal's gamma. Until now this task has not been done to the best of our knowledge. This is somewhat surprising since it was suggested in 1954 by Goodman and Kruskal.

THE SYMMETRIC PARTIAL GAMMA COEFFICIENT

It is clear that if we have two ordered polytomies, gamma may be used as a symmetrical measure of association, and it has a clear (PRE) interpretation. If a third ordered polytomy is added as a test variables, we may calculate three bivariate gamma coefficients, γ_{xy}, γ_{xz}, and γ_{yz}, all of which are symmetrical. Although the gamma coefficient is not ordinarily thought of as indicating the extent of linear association between variables, it may be considered as a general index of monotonicity of the underlying relationship. This being the case, gamma may indicate the tendency of the underlying rank orders that are to be related in a monotonic fashion (5). Gamma indicates the general tendency toward monotonicity.[1] Thus, if we desire a symmetrical partial gamma coefficient, the following measure is proposed:

$$\gamma_{xy.z} = \frac{\gamma_{xy} - (\gamma_{xz} \ \gamma_{yz})}{(1 - \gamma_{xz}^2) \ (1 - \gamma_{yz}^2)} \qquad (3)$$

This formula may be expanded. The second order partial coefficient, where two test variables are used, would be:

$$\gamma_{xy.zw} = \frac{\gamma_{xy.z} - \gamma_{xw.z} \ \gamma_{xy.w}}{(1 - \gamma_{xw.z}^2) \ (1 - \gamma_{xy.w}^2)} \qquad (4)$$

Since the proposed partial is symmetric, it does not necessitate assumptions concerning causal order or time sequence. This measure also takes into account the general monotonic tendencies of all the bivariate relationships. Unlike Davis' (3) asymmetric partial gamma coefficients, this measure is not affected by extended distributions of ordered categories. The proposed coefficient indicates the association between x and y adjusting both x and y for monotonicity on z. The monotonicity tendency between x and y adjusted for their monotonicity with z is represented in the relationships of the residuals when each variable is predicted from z. The symmetric partial association, then, is an association between errors in prediction (in terms of their original association). It should be noted that this partial coefficient is an indicator of monotonicity between x and y adjusting for z only if the initial bivariate relationship is monotonic. This partial coefficient is not appropriate if the zero order relationship between x and y is not monotonic.

AN EXAMPLE

We will use the data provided by Davis (3:192) to illustrate the differences in asymmetric and symmetric partial coefficients. Table 1 shows the distribution of three variables, i.e., age, education, and reading. The intent is to determine the degree of association between age and reading when the effects of education are partialled out.

The bivariate gamma coefficients are:

$$\gamma_{age.\ reading} = -.241$$

$$\gamma_{age.\ education} = -.416$$

$$\gamma_{education.\ reading} = .689$$

From the above coefficients we can compute the symmetric coefficient, which will indicate the degree of association between age and reading when the effect of education is partialled out. Substituting the gamma values in equation (3), we have:

$$\gamma_{ar.e} = \frac{-.241-\ (.689)\ (-.416)\}}{\{1-(.689)^2\}\ \{1-(-.416)^2\}} = -.067$$

The symmetric partial coefficient is slightly higher than the asymmetric partial coefficient (-.014) but essentially does not alter the interpretation suggested by Davis (that there is negligible association between age and reading). Similarly, we can examine the relationship between education and reading when the effect of age is partialled out. From the associations above, we can write:

$$\gamma_{er.a} = \frac{.689-\{\ (-.241)\ (-.416)\}}{\{1-(-.241)^2\}\ \{1-(-.416)^2\}} = .667$$

The asymmetric partials are not computed by Davis in his paper but can be computed easily from his data. The symmetric partial coefficient is .667 indicating that the original relationship be-

TABLE 1

AGE, EDUCATIONAL ATTAINMENT, AND BOOK READING IN A SAMPLE OF BALTIMORE WOMEN*

Education		Age		Book Reading: Low (-)	Book Reading: High (+)
College	(+)	45 or older	(+)	36	101
		Under 45	(-)	46	163
High School		45 or older	(+)	179	159
		Under 45	(-)	327	290
Less than high school	(-)	45 or older	(+)	335	54
		Under 45	(-)	133	24
TOTAL		45 or older	(+)	550	317
		Under 45	(-)	506	477

*Source: Davis (3)

tween education and reading holds regardless of age.

DISCUSSION

The gamma coefficient proposed by Goodman and Kruskal (4) has proven to be of considerable value to behavioral scientists. From this basic measure others have been developed. Somers (9) developed an asymmetric measure based upon the logic of gamma, which adjusted the gamma coefficient for ties. In 1967 Davis (3) developed an asymmetric partial for gamma. Morris (13) has explicated several ordinal measures of multiple correlation, among them gamma and gamma k. Others who have contributed to the development of ordinal measures of association, based upon the logic of gamma, are Leik (7), Leik and Gove (8), Kim (10), and Hawkes (11).

Until now a symmetrical partial coefficient for gamma has not been explicated. It should be noted that the proposed coefficient is similar to that proposed by Kendall (12) for his coefficient Tau-b. Kendall apparently was not aware of the reasons his partial coefficient was so much like that of the product-moment correlation coefficient. By showing how both variance and covariance can be estimated for ordinal data, Hawkes (11) has shown how several ordinal measures of association (including gamma) are analogs of product-moment correlation coefficient. A symmetrical partial gamma should prove useful in situations when we have ordinal data and where causal linkages or time order is not clear.

Although the sampling distribution for partial gamma is not yet known, it is known for zero-ordered gamma (Goodman and Kruskal, [14]). From the zero-order distribution it should be possible to generate a partial gamma sampling distribution.

FOOTNOTE

1. From Bishir and Drewes (1) and McGinnis (6) we know that monotone convergence implies that boundness is tantamount to convergence. If an increasing sequence (x_n) is bounded, it has a least upper bound u. Then u is an upper bound, but for any $r > 0$, the number $u - r$ is not an upper bound. Thus, for a member x_k:

$$u - r < x_k \leq u.$$

Since this is an increasing sequence, we have:

$$u - r < x_k \leq s_n \leq u \text{ for all } n \geq k.$$

This implies that $u = \lim (x_n)$. We should however keep in mind that when we deal with ordinal level measurements, we do not have strict monotones. By strict, we mean that the exact location of inequalities are within an interval, $I|a,b|$.

REFERENCES

1. Bishir, John W. and Donald W. Drewes. Mathematics in the Behavioral and Social Sciences. New York: Harcourt, Brace and World, Inc., Pp. 112-113.
2. Costner, Herbert L. Criteria for Measures of Association. American Sociological Review, 1965, 30, 341-353.
3. Davis, James A. A Partial Coefficient for Goodman and Kruskal's Gamma. Journal of the American Statistical Association, 1967, 62, 189-193.
4. Goodman, Leo A. and William H. Kruskal. Measures of Association for Cross Classifications. Journal of the American Statistical Association, 1954, 49, 732-764.
5. Hays, William L. Statistics for Psychologist. New York: Holt, Rinehart and Winston, 1964. Pp. 641-643.
6. McGinnis, Robert. Mathematical Foundations for Social Analysis. Indianapolis: Bobbs-Merrill, 1965. Pp. 248-250.
7. Leik, Robert K. A Measure of Association for Ordinal Variables. Paper read at the meeting of the American Sociological Association, 1966.
8. Leik, Robert K., and Walter R. Gove. The Conception and Measurement of Asymmetric Monotonic Relationships in Sociology. American Sociological Review, 1969, 74, 696-709.
9. Somers, Robert H. A New Asymmetric Measure of Association for Ordinal Variables. American Sociological Review, 1962, 27, 799-811.
10. Kim, Jae-On. Predictive Measures of Ordinal Association. American Journal of Sociology, 1971, 76, 891-907.
11. Hawkes, Roland K. The Multivariate Analysis of Ordinal Measures. American Journal of Sociology, 1971, 76, 908-926.
12. Kendall, Maurice G. Rank Correlation Methods. New York: Hafner, 1948.
13. Morris, Raymond N. Multiple Correlation and Ordinal Scaled Data. Social Forces, 1970, 48, 299-311.
14. Goodman, Leo A. and William H. Kruskal. Measures of Association from Cross-Classifications, III, Approximate Sample Theory. Journal of the American Statistical Association, 1963, 58, 319-364.

CONSUMER ATTITUDES TOWARD AUTOMOBILE SAFETY MEASURES: A CLUSTER ANALYTIC APPROACH

Frederick Wiseman, Robert Lieb, Mark Moriarty
Northeastern University

1. Introduction

Automobile safety programs, which range from law enforcement to structural design of automobiles, influence not only the safety of the driving public, but also their expenditures. Because consumers are so directly affected by such programs which are implemented by various governmental units, often in conjunction with automobile manufacturers, a national probability survey was conducted by Lieb and Wiseman in January, 1973, to determine consumer attitudes toward a number of existing and proposed automobile safety programs.[1] Questionnaires were mailed to 888 households and after one follow-up postcard, 420 (47 percent) usable replies were received.

The survey results indicated general public support for existing and proposed automobile safety programs. Evidence of this was provided by the following findings:

1. Seventy-seven percent of those respondents who lived in states which had mandatory automobile inspection programs believed that their state's program was effective in promoting automobile safety.

2. Eighty-six percent of the respondents called for a continuation of the leading role played by the federal Department of Transportation in the development of automobile safety programs.

3. While seventy percent of the respondents were not in favor of a proposed 1976 requirement for inclusion of air bags in new automobiles, they were evenly divided as to whether they would purchase an optional air bag at a price of $100.[2]

4. Fifty percent of those surveyed indicated that they would purchase an optional $750 protection package which would make their new automobile "fatality proof."

5. Seventy-two percent of respondents supported a proposed 1975 federal regulation which would require the inclusion of speed-control devices (governors) in new automobiles. Such devices would prevent new automobiles from traveling in excess of 95 mph.

6. Respondents also favored severe penalties for persons convicted of drunken driving offenses. Evidence of this was provided by the fact that 26% of the respondents recommended prison terms for individuals convicted of drunken driving in accidents which resulted in only property damage. This percentage jumped to 56% when the offense involved non-fatal personal injury and to 83% when the accident resulted in fatal injuries.

While the interest of the public in automobile safety was evidenced by a relatively high rate of response to the survey, it was ironic that many respondents failed to take full advantage of existing automobile safety equipment (respondents used seat belts approximately 50% of the time.)

This paper presents the results of subsequent research which was conducted for the purpose of determining whether respondents could be grouped into a small number of homogeneous segments on the basis of their similarity of attitudes and opinions toward automobile safety related issues. The technique used in this effort was cluster analysis. The specific computer program utilized is discussed in the following section.

2. METHODOLOGY

Cluster analysis is a general term for a large class of numerical procedures whose purpose is to define groups of objects which are related to each other based on some measure of proximity. Most variants of the procedure attempt to develop high within group homogeneity and among group heterogeneity in terms of the proximity measure. The proximities between objects may be matching coefficients, correlation coefficients, distance measures, or anyone of a number of other measures, the only requirement being that a rank order of pairwise proximities obtains.

The Johnson Cluster Program used in the analysis is a nonmetric, hierarchical algorithm and the measure of respondent similarity is the correlation coefficient between respondents computed across subjects' questionnaire responses in the sample. Note that instead of correlating variables across sample respondents, the interest is in similarity of respondents or the resemblance of subject profiles across a set of variables. That is, respondents are judged as similar or dissimilar to each other based on the magnitude of the correlation coefficient between them computed on the basis of their responses to the questionnaire.

The Johnson algorithm is hierarchical in that at the initial stage, each subject is his own cluster and at each succeeding stage, the program chooses that pair of subjects which are most similar, predicated on the rank order of proximities. The program adds a new subject to a cluster if the minimum value of its correlations with each of the existing cluster members exceeds the correlation of any two unclustered subjects. If this is not the case, a new cluster of two subjects is formed.

For the Johnson Cluster program, subjects, once having entered a cluster, remain and, as the algorithm proceeds, all clusters merge until one large cluster exists. While no objective rules exists for deciding, on the basis of statistical tests, when the process should terminate, it is inappropriate to consider doing so any way. The implicit a priori assumption is that clusters exist in the first place, and further, the procedure is exploratory in the sense that the clusters are formed from the data and not from objective external presumptions. While rules of thumb are dangerous to suggest since the research questions may be unique in any given context, rules based on minimum tolerable and meaning cluster size seem most appropriate.

Due to the limitations of the clustering program, the entire sample could not be used, but rather only a subsample of 200 respondents. It is with this subsample that the analysis to be reported on in this paper is based.

3. CLUSTERING VARIABLES

Five variables (CV_1, CV_2, CV_3, CV_4, CV_5) were used in order to group survey respondents into homogeneous groups. Each clustering variable was selected to represent a major automobile safety related issue. The variables were a respondent's

CV_1: Opinion as to the nature of the role that should be played by the DOT in setting auto mobile safety standards (1=Major role,...., 5= no role);

CV_2: Current percentage rate of seatbelt utilization;

CV_3: Likelihood of purchasing air bags as optional equipment at a cost of $100 (1=definitely yes,...., 5=definitely no);

CV_4: Likelihood of purchasing an all encompassing $750 total protection package (1= definitely yes,..., 5=definitely no); and

CV_5: Opinion as to the severity of penalties that should be given to convicted drunken drivers (1=light,..., 5=harsh)

Data were also obtained on six supplementary variables (SV) for each respondent

SV_1: Opinion as to whether air bags should be made optional equipment (yes=1; no=0)

SV_2: Opinion as to whether governors should be made mandatory equipment (yes=1; no=0)

SV_3: Current percentage rate of shoulder belt utilization

SV_4: Age (1=under 25; 2=25-34; 3=35-54; 4=55 and over)

SV_5: Income (1=under $7,500; 2=$7,500-$15,000; 3=over $15,000)

SV_6: Education (1=attended grade school; 2=attended high school; 3=high school graduate; 4=attended college; 5=college graduate; 6=graduate degree)

The above supplementary variables were not used in the formation of the clusters, but were used to aid in their description.

4. RESULTS

The clustering procedure produced a total of 8 clusters at a correlation level of .55. That is, on a respondent by respondent basis, responses of each member of a particular cluster correlated at least at the .55 level with the responses of each additional member of the cluster. Out of the 200 total sample members, only 21 (10.5%) did not fall into one of the eight clusters. The mean values for each cluster on each of the five clustering and six supplementary variables, together with the size of each of the groups are presented in Table 1.

From the data in Table 1, the clusters can be described as follows:

Cluster 1 members are opposed to the present role being played by the DOT in setting mandatory automobile safety standards. Further, these individuals show little interest in new safety equipment and all believe that airbags should be made optional rather than required equipment. However, they do believe in the issuance of stiff penalties to those individuals found guilty of drunken driving violations.

Cluster 2 is the most extreme of all clusters as virtually no interest in safety programs is expressed by group members. They do not wear their safety belts and do not plan to purchase airbags or the $750 total protection package. In addition, this cluster indicates the greatest opposition to the present role being played by the DOT and are by far the most lenient of all groups in the area of what penalties should be assessed to drunken drivers. Further, members are substantially older and somewhat less educated than are individuals in most other clusters.

Cluster 3 individuals are in favor of almost any governmental automobile safety program. Virtually all indicate plans to purchase the protection package and airbags if they are made optional equipment. Members are also among the most frequent users of safety belts and believe in strict penalties for convicted drunken drivers.

Cluster 4 members are quite similar to Cluster 3 members except in the area of safety belt utilization. Surprisingly, little use is made of either the seat or shoulder belt. However, all are in favor of airbags being mandatory and all express purchase intentions if they are made optional at a cost of $100.

Cluster 5 contains individuals who give the most severe penalties, usually prison terms to convicted drunken drivers. These people also make frequent use of their safety belts, but, for the most part, are indifferent to other safety measures.

Cluster 6 is similar to Cluster 5 except in the area of safety belt utilization. This segment makes infrequent use of their seat and shoulder belts.

Cluster 7 is also similar to Cluster 5 except that members are much more lenient in the area of drunken driving penalties.

TABLE I

Cluster Means

Variable	1 (n=10)	2 (n=9)	3 (n=10)	4 (n=10)	5 (n=29)	6 (n=14)	7 (n=32)	8 (n=57)
Clustering Variable								
Role of DOT	3.2	3.7	1.0	1.0	1.0	1.0	1.0	1.0
Seatbelt utilization	60%	5%	98%	45%	93%	25%	83%	25%
Airbag installation	3.3	4.4	1.0	1.0	3.2	3.2	3.4	3.2
Protection package	2.8	3.7	1.4	1.4	1.8	2.0	1.8	1.7
Penalties	4.1	1.9	3.9	3.6	4.5	4.1	2.8	2.6
Supplementary Variable								
Airbags optional	1.0	.8	.6	.2	.7	.7	.7	.8
Governors	.7	.4	.7	1.0	.7	.8	.8	.8
Shoulderbelt utilization	10%	3%	40%	13%	33%	10%	28%	8%
Age	2.7	3.7	2.8	2.6	2.7	2.7	2.8	2.6
Income	2.4	2.2	2.3	2.5	2.1	2.0	2.3	2.1
Education	4.3	3.8	4.3	3.8	4.3	4.0	3.8	3.7

TABLE II

Results of Canonical Analysis

Set 1: Group Membership Variables		Coefficient
	X_1	.595
	X_2	.771
	X_3	-.079
	X_4	-.038
	X_5	-.130
	X_6	-.031
	X_7	-.070
Set 2: Clustering Variables		**Coefficient**
CV_1	Role of DOT	.952
CV_2	Seatbelt Utilization	.104
CV_3	Airbag installation	.006
CV_4	Protection package	.030
CV_5	Penalties	-.072

Canonical Correlation	.95
Willes Lambda	.008
Chi Square	789.93
Degrees of Freedom	35
Significance level	.00

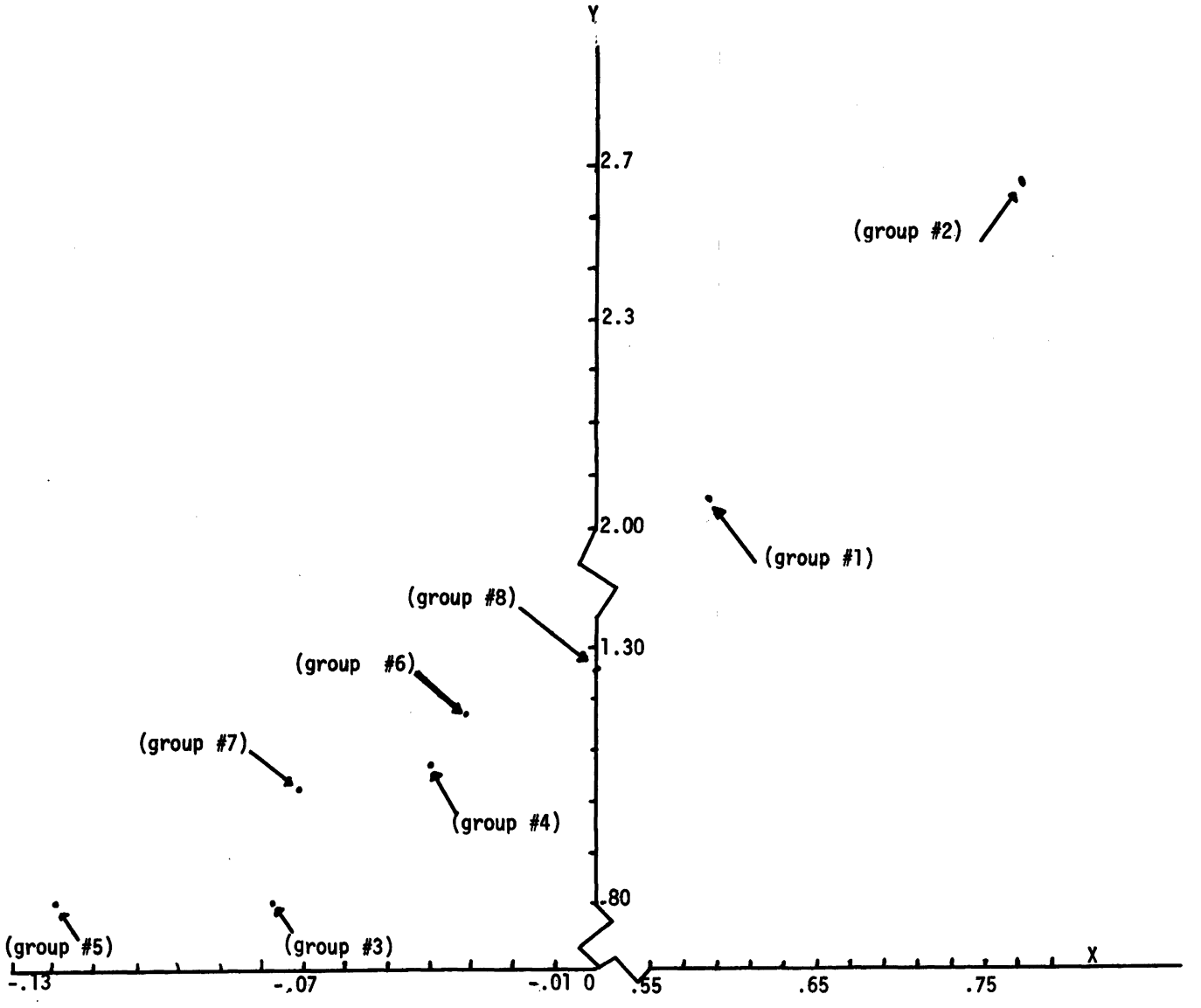


FIGURE 1. Canonical Analysis Plot

Cluster 8 is the largest of all clusters and shows no extreme viewpoints. All, however, believe that the DOT should play a major role in the area of automobile safety. It is interesting to note that this segment makes infrequent use of both seat belts and shoulder belts.

A canonical correlation analysis was performed in order to position the clusters in 2 dimensional space. The two sets of variables used in the analysis were, first, the seven dummy (0,1) variables necessary to indicate cluster membership and, second, the five previously discussed clustering variables. To represent cluster i, (i=1, 2,...7), one takes the canonical coefficient of variable x; as the X co-ordinate and $\sum_j B_j CV_{ji}$ as the Y co-ordinate where:

B_j is the canonical coefficient of the j^{th} clustering variable and

$\overline{CV}_{ji}$ is the mean value for cluster i on the j^{th} clustering variable

Due to the nature of the dummy variables, the X coordinate for cluster 8 will be zero. The results of the canonical analysis and the plot are given in Table 2 and Figure 1, respectively.

As can be seen, the X dimension in Figure 1 appears to be a measure of respondent's opinion as to what role the DOT should play in setting standards. The Y dimension appears to represent the amount of protection that drivers want on the road. Note that clusters 1 and 2 are the most extreme clusters and combined represent about 10% of the population.

CONCLUSIONS

The above analysis has grouped respondents on the basis of similarity of attitudes and opinions toward a number of automobile safety related issues, rather than on socio-economic and demographic profiles. The results indicate that there is considerable heterogeneity across groupings. While there are small extreme clusters holding either strong positive or negative opinions toward governmental automobile safety proposals, the majority of individuals are clustered in groups that have ambivalent attitudes toward a number of these issues.

This lack of public commitment is especially obvious with respect to the airbag and protection package issues raised in this study. In view of this apparent ambivalence, it appears that public opinion may be swayed to support such proposals. If the DOT believes that widespread public support is necessary to bring about future legislation in these areas, it should consider expansion of its public education efforts.

FOOTNOTES

[1]Robert C. Lieb and Frederick Wiseman, "Consumer Attitudes Toward Automobile Safety Programs," Technical Paper 73-ICT-36, presented at the Intersociety Conference on Transportation, Denver, Colorado, September, 1973.

[2]An airbag is a balloon-like device which inflates in the automobile's passenger compartment in the event ofhead-on collisions. The bag is designed to act as a cushion between occupants and the instrument panel. Following impact, the airbag immediately begins to deflate.

MINUTES OF THE ANNUAL MEETING OF THE SOCIAL STATISTICS SECTION

New York, New York, December 29, 1973

The meeting was opened by Theodore Woolsey, Chairman, at 5:15 p.m. As a result of the recent election, the officers for 1974 are:

Chairman	Charles B. Nam
Chairman-Elect	Joseph Waksberg
Vice Chairman (1973-74) (Program Chairman)	Denis F. Johnston
Vice Chairman (1974-75) (Asst. Program Chairman)	Mollie Orshansky
Secretary (1974-75)	Mary G. Powers
Representative on the National Board of Directors (1973-75)	Evelyn M. Kitagawa
Representative on the National Council (1974-75)	Eva L. Mueller
Publications Liaison Officer (1973-74)	Monroe Lerner

Edwin D. Goldfield was appointed Editor of the Proceedings for the 1973 meeting.

The representative on the National Council, Eli Marks, gave a report on the recent joint meeting of the Council and the Board. (1) The revised constitution will be submitted to the members about April 1974. The Board will be enlarged and will assume many functions previously done by the Council. Each section will have two representatives to the Council and one or more representatives to the program committee. The second Vice-Chairman will be a member of the program committee. (2) Resolutions on the status of women in the American Statistical Association were passed. (3) The Board approved further work by the ASA Conference on Surveys of Human Populations. A research staff, with support by the National Science Foundation, has been proposed to investigate causes of difficulties in doing surveys and methods to improve them.

The Publications Liaison Officer, Monroe Lerner, distributed and discussed an announcement asking for expository papers for The American Statistician.

The Program Chairman for 1973, Daniel Horvitz, reported that the program had a good balance of topics. He suggested that all papers in the future have some emphasis on methodology rather than reporting only substantive findings. A vote of thanks was given to Dr. Horvitz for doing a fine job and for accepting the position at the last minute.

Some members disagreed with the recommendation that the program should be so heavily methodological. The history of the Section should be studied to determine its purposes. The Section should decide on what it should do. It was agreed that the program for 1974 would be a balance of methodology and subject matter, and that there should be further discussion of this topic at later meetings.

The Program Chairman for 1974, Denis Johnston, said that the theme for invited papers sessions would be social indicators. Detailed plans for many of the sessions were presented and discussed. It was suggested that a few meetings should be on topics other than indicators. One proposed topic was sources and uses of data; for example, the use of census data by community groups. It was requested that suggestions for additional topics be sent to the program chairman.

Procedures for nominations of Fellows of ASA were called haphazard, resulting in late or no nominations of qualified persons. A committee of the Washington, D. C. Chapter of ASA reviews lists of their members, finds persons who should be but are not Fellows, and arranges for persons to nominate them. It was suggested that the 1974 officers of this Section consider the appointment of a committee to work on nominations of qualified members of the Section.

A motion was passed to include in the Proceedings as an addendum to these minutes the report of the Section for 1973 that was prepared by Mr. Woolsey.

Mr. Woolsey reported on the value of the meeting of the Section's officers in February, and urged that such a meeting be repeated in future years. The problem of travel expenses is difficult but often can be solved by using ingenuity, he believed.

The attached annual report describes the assignment to Monroe Lerner re social indicators. Dr. Lerner said that there is not now a need for an advisory committee of the Section. It had been suggested that the Section might set professional standards with regard to statistics for social indicators. In the discussion it was brought out that many persons are now working on this topic and that the previous need for work by the Section on social indicators no longer exists. The following motion was then made, seconded, and passed:

> The Social Statistics Section should report to the American Statistical Association that it does not see a need for an advisory committee on social indicators at this time. This recommendation is based on several years of consideration of Recommendation 40 of "A Study of Future Goals of ASA", 1971.

The letter to ASA about this motion should give reasons for the motion.

A request for a subsection on Survey Research Methodology was received from The Survey Research Interest Group. The following motion was made, seconded, and passed:

The Social Statistics Section supports the establishment of a Subsection on Survey Research Methodology.

This subsection would arrange programs about survey research methodology, and the Social Statistics Section would organize sessions on other topics. The need for a mail ballot of the Section's members about this motion will be determined.

The meeting closed at 7 p.m. after a vote of thanks for the fine work by Theodore Woolsey as chairman in 1973.

1973 Officers of the Social Statistics Section

Chairman	Theodore D. Woolsey
Chairman-Elect	Charles B. Nam
Vice-Chairman (Program Chairman)	Daniel G. Horvitz
Vice-Chairman	Denis F. Johnston
Secretary	Regina Loewenstein
Representative on the National Board of Directors	Evelyn Kitagawa
Representative on the National Council	Eli S. Marks
Publications Liaison Officer	Monroe Lerner
Editor of Proceedings	Edwin D. Goldfield

REPORT OF THE SOCIAL STATISTICS SECTION, ASA, FOR THE CALENDAR YEAR 1973

First, it should be indicated that the "routine" business of the Section has proceeded reasonably smoothly. In the fall of 1972, the senior Vice-Chairman of the Section, who was to have been in charge of the Program for 1973, resigned the office. With the approval of the Chairman of the Section, the Chairman-Elect found a substitute to complete the 2-year term. This was Dr. Dan Horvitz who was at that time with the Research Triangle Institute and is now on the faculty of the University of North Carolina. He was duly appointed and has worked hard. With the help of others he has produced what appears, at this writing, to be an outstanding program for the meetings in New York City.

A Nominating Committee, consisting of W. Parker Mauldin, Chairman; Thomas Jabine; and Otis Dudley Duncan, was appointed to nominate candidates for the posts of: Chairman-Elect for 1974; Vice-Chairman for 1974-75 to serve as Assistant Program Chairman in 1974 and Program Chairman in 1975; Secretary; and Section Representative on the Council for 1974-75. The nominations were submitted to headquarters on April 20, 1973, and Joseph Waksberg, Mollie Orshansky, Mary Powers, and Eva Mueller, respectively, were elected.

Of a non-routine nature was the holding of a mid-winter, one-day meeting (Feb. 27, 1973) of the officers of the Social Statistics Section. ASA headquarters made the meeting room available, and travel costs were covered by a variety of underhanded strategems. (One that was not underhanded was to schedule the meeting the day after the meeting of the Board so that our representative on the Board could have her travel paid by ASA).

We feel that the meeting was extremely helpful in providing some continuity to the affairs of the Section and strongly recommend that some way be found to permit such mid-winter meetings to be held routinely. For example, it gave all the officers an opportunity to comment upon the plans being formulated for the Program, to hear reports from our representatives to the Council, the Board, and the Publications Committee, and to discuss possible committees of the Section. The Annual Business Meeting will also be far more useful because of this opportunity to review the business of the Section. It is the only opportunity for the Officers of the Section to meet face-to-face except at the annual meeting. Dr. Leone met with us a part of the day and Mr. Bisgyer the whole day.

A particular agenda item at the February 27 meeting was a piece of business left over from the previous year, that is, the plan for a Section Committee on Social Indicators. Since the outcome of that discussion has been passed on to Dr. Leone in our Section's report on the status of recommendations from the Future Goals Report, it will not be reported here. Suffice it to say that because of much skepticism from all hands (including people asked to be chairman of such a committee but declining) as to whether the committee could find something to do that wasn't already being done, the Section has asked Dr. Monroe Lerner to make certain specific explorations and report on them at the December business meeting. At that time a final decision should be made about whether to establish a Committee with a limited and well-defined charge or report back to the Board of Directors that the Section is not in favor of the idea.

Theodore D. Woolsey, Chairman