

PROBLEMS OF NONSAMPLING ERROR IN THE
SURVEY OF INCOME AND EDUCATION: CONTENT ANALYSIS

Robert E. Fay III, U.S. Bureau of the Census

Introduction

Congress mandated the 1976 Survey of Income and Education (SIE) through the legislative injunction to "expand the current population survey (or make such other survey)" to furnish current estimates by State of the number of school age children living in poverty families. The legislation directed that these estimates be analyzed for possible use in the allocation of educational monies to school districts under Title I of the Elementary and Secondary Education Act of 1965. The Congress further enjoined the Secretaries of Commerce and HEW to submit a report on the survey, "including analysis of its accuracy and the potential utility of the data derived therefrom..." for updating this allocation. By agreement between the two departments, the Bureau of the Census assumed responsibility for the analysis of the accuracy of the survey results and the consequent direct implications for the question of utility.

This paper will describe the design and underlying principles of the Census Bureau's evaluation program for the SIE. Because the report on the analysis is currently under review and revision and has not yet been submitted to the Congress, it is inappropriate to discuss publicly the estimates or conclusions of the evaluation program out of respect for the Congress. This paper will, however, present a statistical model that forms a component of the analysis, since this model is based entirely upon published data.

Considerations in the Design

The current government definition of poverty for statistical purposes is based principally upon money income and number of persons in the family, although the age and sex of the head, and the farm/non-farm status of the household are also included in the determination. Previous experience, particularly from the comparison of the Census with the Current Population Survey in 1970, has shown that the statistical measurement of poverty at the national level is sensitive to the choice of survey procedures. Furthermore, although differences in coverages and in definitions of household membership may contribute to differences between surveys, the available evidence pointed to problems in the collection of income data and in allocation for non-response as the primary driving force behind these differences. In general, obtaining accurate and complete income data from surveys and censuses has been problematic, but this difficulty has appeared most conspicuously in the poverty statistics.

These considerations suggested the particular formulation of the accuracy of the estimates in terms of their consistency among States. In other words, because the primary impetus for the survey was to obtain current estimates for use in an allocation formula, the survey results would be accurate for this purpose if they led to a

correct allocation among States. As a first approximation, this would in turn be achieved by survey estimates that correctly represented each State's share of the national total of children aged 5-17 in poverty families, even if the national total was open to question.

Discussions between the Executive Branch and Congress led to the agreement for a specification of a coefficient of variation of 10 percent for each State's estimate of the number of children aged 5-17 in poverty families. Although the relation between this objective and the actual statistical reliability obtained by the SIE is an important question, the primary focus of the evaluation was to determine the possible effect of non-sampling errors in the State estimates.

Several previous evaluation programs to measure non-sampling error have been formulated in terms of the consistency of the respondents' answers over repetitions of the survey process or the uniformity of the interpretation and execution among interviewers. In planning the SIE evaluation, the analytic measures obtained from these other studies, "simple response variance" and "correlated" or "interviewer variance," were seen as at best tangentially related to the problem of non-sampling error in the SIE State estimates based upon the work of many interviewers and thousands of interviews. The perspective chosen instead was to determine directly the presence of systematic non-sampling errors affecting the SIE State estimates. This perspective led in turn to the decision to create an alternative survey process as a standard for comparison to the SIE. By conducting an alternative process of greater intensity than the SIE, the SIE survey estimates would be judged consistent within the limits of this standard if the more intense procedures would not change the allocation among States. Variance considerations forced this evaluation to be a reinterview of a subsample of the original sample, but conceptually the principles of analysis would have been similar if an entirely independent (but necessarily larger) evaluation survey had been conducted.

Because of an increasing legislative tendency to distribute public monies to subnational units according to need and to measure this need statistically, an increasing obligation has been placed upon the producers of these statistics to insure the consistency of the measurement process. The conceptual design for this evaluation may therefore serve as an example for future evaluations of this sort.

Design of the Reinterview

To create a standard for the evaluation of the SIE, two principles were followed: to obtain the critical information in the households selected for reinterview as independently as possible, and to increase the intensity of effort sufficiently to establish prime facie evidence that the

reinterview was indeed a valid standard for evaluation. As a consequence, the planning for the reinterview required an effort comparable to the planning for a new survey.

In the SIE and CPS, generally one person in a household served as the "household respondent" and provided all information, including on income, for all household members. A first specification in the reinterview design was to require self-response for all household members age 16 and over, even though call-backs were generally necessary to achieve this. Although hypothetical situations can be constructed where a self-respondent is less informed or cooperative than another household member, in general self-response was felt to be a better, although expensive, choice for the reinterview. (Some Census Bureau surveys, notably the National Crime Surveys, have required self-response when it has been judged that an increase in accuracy would justify a concomitant increase in cost.)

The second key feature of the design was the development of a new questionnaire. The new questionnaire incorporated a deliberate attempt to correct possible deficiencies of the SIE income section, which in turn had represented a minor modification of the corresponding CPS section. The CPS and SIE income sections record the data on reciprocity and amounts in a FOSDIC-readable format. The questions on reciprocity and amounts for each type of income follow each other in alternation. Although indicating all the necessary information to be obtained, this design provides neither the interviewer nor the respondent support in correctly determining amounts. It has also been suggested that the rapid succession of questions on reciprocity leads respondents to say "no" more or less automatically, even when one of the types may actually have been received. Some CPS interviewers have also suggested asking all questions on reciprocity first before any questions on amounts, since the latter are often the most sensitive issues. This would follow a general principle of questionnaire design, to precede the most sensitive questions with less sensitive ones.

The broad structure of the reinterview questionnaire is to establish in a "screening" section the reciprocity by type of income in the context of a general review of possible income-related activities and situations during the year, and then to collect the amounts of the various types of income in "amounts" sections specifically designed for the particular types of income. For example, a respondent is asked in the screening section about all jobs held during the year. Later, in the amounts section for wages and salary, the respondent is questioned on each job separately. For each job, the respondent is first requested to consult a W-2 form for the information but, if unable or unwilling to do that, is allowed to provide an estimate if the respondent feels reasonably certain of the amount. If no figure can be obtained in this way, several alternative paths of questions assist the respondent in constructing an estimate based on an annual salary, an hourly or daily wage, or average amount paid in each paycheck. Consequently, the

reinterviewer is directed through a series of questions that in general a resourceful interviewer might use with respondents requiring such help, but which is not provided by the CPS or SIE questionnaires.

A subsample of about 4.5 percent was selected for reinterview from the CPS and SIE samples. Approximately 2,000 and 6,000 reinterviews were obtained for the CPS and SIE, respectively. Stratification on the number of children aged 5-17 and the originally reported income was employed to reduce the sampling variance of the reinterview estimate of the number of children aged 5-17 in poverty families. In general, reinterviewers were provided only the information required to locate the original household, to insure the independence of the reinterview information. Subsequent edits in the field offices and later by computer identified a group of cases with significant discrepancies. This group was recontacted to assure the accuracy of the reinterview results. The preliminary analysis of these data has been completed. Because they form the basis for the evaluation to be reported to Congress, however, it is fitting to postpone the public discussion of the findings.

A Statistical Model for Children in Poverty

A standard statistical technique, linear regression, illustrates important aspects of the SIE estimates of children aged 5-17 in poverty families by State. Recent work in the application of this technique to survey estimates is due to Eugene Ericksen (1973, 1974). In general terms, sample estimates for the geographic units (counties, SMSA's or States) may be used as the dependent variable in a linear regression based upon symptomatic data gathered without sampling error for the same geographic units. The resulting predicted values are generally biased estimates of the population values for these geographic units, but in some applications they may possess considerably smaller average mean square errors than the sample estimates themselves. Furthermore, this technique allows the linear relationship between the symptomatic variables and the variable of interest to be determined directly from the current sample data, rather than from a priori reasoning or previous experience.

On the basis of this research, Census Bureau staff (Gordon Green and Robert Fay) studied the possible application of this technique to estimate the proportion of children aged 5-17 in poverty families for each State. The model was developed (in 1975) by attempting to fit the 1970 Census values for the percent of families in poverty by State on the basis of the corresponding 1960 Census results and other information. (Estimates of children aged 5-17 in poverty by State are not available from the 1960 Census.) These investigations favored a model based upon the census values and estimates of Per Capita Income (PCI) published in the Survey of Current Business by the Bureau of Economic Analysis. Sample estimates for the percent of children in poverty by State are fitted by a regression incorporating six independent variables: the constant term, the

census percent in poverty for the base year, and two variables derived from PCI figures for each of the base and current years. For each of two years of BEA data, the median, PCI_m , of the 51 State figures is determined and the variables

$$X_{j1} = \begin{cases} \ln(PCI_j / PCI_m) & \text{if } PCI_j > PCI_m \\ 0 & \text{otherwise} \end{cases}$$

$$X_{j2} = \begin{cases} 0 & \text{if } PCI_j > PCI_m \\ \ln(PCI_j / PCI_m) & \text{otherwise} \end{cases}$$

formed. The regression is weighted inversely proportional to the sampling variance of the sample estimates.

Ericksen's research included a possible approach to estimate the average mean squared error of the regression estimates. Basically, the sampling error of the sample estimates may be subtracted from the squared deviations between the sampled and fitted values to estimate the squared bias of the regression as the remainder. In this way, a current evaluation of the regression estimates may be obtained. The technique generally requires precise estimates of the sampling errors, however, and becomes ineffective in cases where the sampling errors completely dominate the biases of the regression.

Although the regression model has been fitted to the sample estimates of the percent of children 5-17 in poverty families by State from the CPS for all years subsequent to 1970, the sampling errors of the CPS estimates obviate any effective assessment of the fit. The SIE therefore affords the first such opportunity since the 1970 Census. Table 1 compares the 1970 Census estimates for 1969, the 1976 SIE estimates for 1975, and the model estimates based upon the SIE by State. The national poverty rates from the SIE and 1970 Census are virtually the same (14.5 percent vs. 14.8 percent), but there is a substantial redistribution of poverty among States. The changes estimated by the SIE since the 1970 Census correspond to an average of approximately 23 percent root mean square (r.m.s.) by State. Since the SIE estimates have an average c.v. of 10 percent, a real change of approximately 20 percent (r.m.s.) may be inferred ($23^2 \doteq 20^2 + 10^2$). On the other hand, the model estimates are within about 14 percent (r.m.s.) of the SIE values, leaving an unexplained bias of only about 10 percent r.m.s. ($14^2 \doteq 10^2 + 10^2$) between the model and SIE estimates. The model estimates therefore describe approximately 75 percent of the real change indicated by the SIE (10 percent r.m.s. vs. 20 percent r.m.s.).

The concurrence between the SIE and regression estimates has two important implications. The result reflects generally well upon the regression methodology: although not free from bias, the results closely resemble the actual survey outcome. Furthermore, if the regression estimates were to continue to explain 75 percent of the real change (a reasonable assumption

according to the original research based on predicting the 1970 Census values from the 1960 Census), while the velocity of real change were also to continue at the rate for 1970-1976, it could be argued that the regression estimates based upon CPS data would be less biased estimates of the actual rates two or three years hence than the SIE rates for 1976.

The logic of the comparison may be reversed, however, and used to argue the face validity of the SIE survey estimates. Linear regression is a projection in the mathematical sense. In the application here, the model estimates are the projection of the 51 survey estimates onto a subspace of dimension 6. The residuals of the regression lie in a subspace of dimension 45. The residual subspace includes most of the sampling error in the SIE survey estimates, as well as the biases of the model estimates. If the SIE State estimates were subject to non-sampling errors, it might be assumed that the largest component of this error would also lie in the residual subspace. Therefore, the 14 percent r.m.s. difference between the model and survey estimates serves as an upper bound on the sum of the sampling error and this component of the non-sampling error. Even though 14 percent may be large, it still provides reassurance that the non-sampling errors of the survey State estimates are not extreme and arbitrary.

References

- Ericksen, Eugene P. (1973), "A Method of Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas," Demography, 10, 137-60.
- _____ (1974), "A Regression Method for Estimating Population Changes for Local Areas," Journal of the American Statistical Association, 69, 867-75.

Table 1. Percent of Children 5-17 Years Old in Poverty Families
According to 1970 Census, SIE, and Regression Model

States by Division	1969 Estimate	1975 Estimates	
	Census	SIE	Regression Model
New England			
Maine.....	14.2	15.3	14.2
New Hampshire.....	7.7	10.3	10.5
Vermont.....	11.4	17.8	11.9
Massachusetts.....	8.4	9.3	10.6
Rhode Island.....	11.0	10.5	11.8
Connecticut.....	7.2	8.4	9.6
Middle Atlantic			
New York.....	12.2	13.1	13.8
New Jersey.....	8.7	11.6	10.2
Pennsylvania.....	10.6	12.6	10.9
East North Central			
Ohio.....	9.8	11.6	11.8
Indiana.....	9.0	9.6	10.8
Illinois.....	10.7	15.1	10.8
Michigan.....	9.1	11.3	11.2
Wisconsin.....	8.7	9.4	9.6
West North Central			
Minnesota.....	9.5	9.1	9.7
Iowa.....	9.8	7.9	8.2
Missouri.....	14.8	14.7	14.8
North Dakota.....	15.7	11.5	10.4
South Dakota.....	18.3	13.1	15.3
Nebraska.....	12.0	10.1	10.3
Kansas.....	11.5	8.6	10.2
South Atlantic			
Delaware.....	12.0	10.4	12.3
Maryland.....	11.5	10.7	11.2
District of Columbia.....	23.2	15.7	17.8
Virginia.....	18.2	13.7	15.0
West Virginia.....	24.3	18.9	18.2
North Carolina.....	24.0	17.8	20.2
South Carolina.....	29.1	23.9	23.4
Georgia.....	24.4	21.3	20.9
Florida.....	18.9	21.6	16.6
East South Central			
Kentucky.....	25.1	21.4	20.2
Tennessee.....	24.8	20.5	20.2
Alabama.....	29.5	15.9	23.1
Mississippi.....	41.5	32.6	32.2
West South Central			
Arkansas.....	31.6	21.4	23.8
Louisiana.....	30.1	22.9	23.8
Oklahoma.....	19.5	14.6	16.2
Texas.....	21.5	20.5	17.7
Mountain			
Montana.....	12.9	12.5	10.8
Idaho.....	12.0	11.0	10.5
Wyoming.....	11.2	8.6	8.2
Colorado.....	12.3	10.7	10.7
New Mexico.....	26.3	26.0	21.2
Arizona.....	17.5	16.8	16.1
Utah.....	10.0	8.0	9.4
Nevada.....	8.8	11.0	9.8
Pacific			
Washington.....	9.3	10.0	10.2
Oregon.....	10.3	8.4	10.2
California.....	12.1	13.8	12.5
Alaska.....	14.6	6.4	6.9
Hawaii.....	9.7	9.6	9.8