PROBLEMS OF NONSAMPLING ERROR IN THE SURVEY OF INCOME AND EDUCATION: COVERAGE EVALUATION

Eli S. Marks and Harold Nisselson, U. S. Bureau of the Census

The primary purpose of the Survey of Income and Education (SIE) was to estimate by state, the number of children aged 5-17 in poverty families. Since the SIE was conceived as an exploration of the feasibility of using intercensal estimates of children in poverty families from a sample survey for allocating Federal aid to education funds, the legislation authorizing the study also contains a requirement that there be an evaluation of the accuracy and utility of the SIE results. Part of this evaluation involves the estimation of nonsampling error effects on the survey estimates.

Apart from sampling error, estimates of the number of children in poverty families are affected by "content errors" in reporting income and age and by "coverage errors" (primarily omissions) in reporting persons and housing units. To study content errors, a subsample of the housing units included in the SIE was selected and reinterviewed. This content evaluation by use of a Reinterview Sample is discussed in another paper presented at this session.[1/]

The SIE Reinterview Sample was also used to estimate coverage error due to the omission of persons in housing units included in the original SIE sample. This also included a check on the coverage of persons in SIE sample households that were erroneously classified as vacant. There were, however, coverage errors due to omission of housing units from the SIE sampling frame. To check on the coverage of housing units (i.e., on housing units omitted from the sampling frame), a coverage check was carried out on a sample of housing units linked to the SIE Reinterview Sample.

Obviously, a sample of housing units to check on the sample frame coverage has to include housing units not in the original sample frame. Ordinarily this involves selecting a sample of areas (segments), listing all the housing units in the sample areas and determining which of the listed housing units are in the sampling frame, i.e., had a chance of being included in the sample. However, the costs and problems of delineating sample areas of satisfactory size for a housing unit coverage check were substantial. It was, therefore, decided (a) to use alternative sampling procedures which did not require delineating (exact) boundaries for sample areas; and (b) to restrict the coverage check to those sections of the population where we would anticipate substantial undercoverage. It was also decided not to check the coverage in the 'New Construction' and 'Special Places' strata of the sampling frame since, for these strata, the difficulties and costs of matching a listed housing unit to the frame would be very considerable and the yield in terms of missed housing units was expected to be small.

The Within Structure Listing Check

For purposes of the SIE coverage check, two classes of missed housing units were defined--(1) missed housing units in enumerated structures (i.e., in structures included in the SIE sampling frame) and (2) housing units in missed structures. Missed housing units in enumerated structures obviously involve multi-unit structures or structures which existed in 1970 and had residential quarters but which have been 'converted' to some other housing unit layout since 1970.[2/] Thus, they involve housing units which existed in 1970 but were missed by the 1970 Census and housing units (or non-housing unit living quarters) created since the 1970 census in structures built prior to the census. The missed units in converted structures are mostly in urban areas (particularly in central cities of SMSA's). The other missed units in enumerated structures are also mostly in urban areas since they involve multi-unit structures. Special problems exist in measuring coverage errors associated with converted enumerated structures since conversion can reduce, increase or leave unchanged the number of housing units in a structure.

To check on missed units in enumerated structures, the structures in which each of the Reinterview Sample (regular) housing units were located were relisted and the relistings for any multi-unit structures (shown as "multi-unit" either in the 1970 Census Address Registers or in the relistings) were matched to the Address Registers and the missed units were identified. Since the original SIE sampling procedure provided for relisting and resampling multi-unit structures where the sampling unit originally selected could not be identified,[3/] structures which had been relisted for the original SIE were omitted from the within structure coverage check. However, this left a substantial number of multi-unit structures in which there was trouble in identifying each of the housing units listed in the Census Address Register with a corresponding unit on the Within Structure Listing form. All such structures were treated as 'converted structures'.[4/] A sample unit (or units) was selected for interview within the 'converted' structure, the effect of "net coverage error" being defined as the difference between the results obtained in the coverage check interview(s) and the results obtained in the original SIE interview.[5/]

Where all the Address Register listings for a multi-unit structure matched units on the Within Structure Listing (WSL) form but there were additional units on the WSL form, these additional units were identified as 'missed housing units' and interviews taken to determine the characteristics of the occupants.

Where a structure that contained an SIE Reinterview Sample unit had more than 12 housing units, it was to be subdivided (by the SIE reinterviewer)

into 'chunks' (floors, wings, etc.) with 12 or less housing units and only one of these 'chunks' (the one containing the original sample unit) was to be listed. Thus, for larger structures, the relisting was of a subsample rather than the entire structure. This does not alter the basic procedure for estimating 'within-structure misses' but merely the specific sampling probabilities involved in making the estimate.

The Successor Structure Check

The SIE sampling procedure departed from that used in the CPS with respect to:

1) Sample selection in multi-unit structures:- Here, CPS relists and resamples multi-unit structures whenever a unit from such a structure is selected for the sample. The SIE relisted and resampled multi-unit structures only when it was not possible to identify the housing unit originally selected for the sample.

2) In rural areas and other areas without clearly identified addresses (street or road names and house numbers), the CPS selects a sample of small areas (segments). The SIE selected individual housing units from the 1970 Census Address Registers (just as in the "address E.D.'s") and the interviewers located these on the basis of whatever information was available (name of 1970 household head, box number, E.D. map spotting).

The Within Structure Listing already described gives an estimate of the effects of the relatively minor modification in the procedure for sampling multi-unit structures. More important from the standpoint of future sampling methodology at the Census Bureau, was a check on the effects of the change in rural areas and small towns from a segment (area) sample to a list sample. To provide a measure of these effects, a Successor (Structure) Check was done. Here, the SIE reinterviewer was instructed to start from the structure containing the reinterview sample unit,[6] (where he reinterviewed the sample unit and completed a Within Structure Listing form); and, proceeding always to the right without crossing a street or road unless it came to a dead end, to list all the structures he encountered until he had listed, in addition to the sample unit, four 'successor' structures built before the 1970 Census. The SIE reinterviewer was to list the names of the current household head(s) and the head(s) in 1970 and any address or description and to check whether the unit was built before or after the 1970 Census.

As shown in Table 1, the 1970 Census housing coverage check[7] had indicated that omissions of entire structures from the Census was most common in rural areas and in urban E.D.'s outside urbanized areas. It also shows a relatively small missed rate for housing units in missed addresses in the larger urban places.[8] For this reason, the Successor Check was restricted to rural areas and to urban places of less than 10,000 population outside of urbanized areas.

In the rural areas and the very small urban areas, 'addresses' are usually not specific to a structure and are, therefore, not of much use for matching. Determination of which successor structures were listed in the 1970 Census Address Registers had to depend on matching names of 1970 household head and map locations of the structures. The map locations were obtained for the Census from the instruction given in 1970 to enumerators in rural areas to draw a small box on the E.D. map to indicate the location of each structure listed in the Address Register, labeling it with the Census Serial Number(s) for the structure. For the Successor Check, the reinterviewers were instructed to draw a sketch map, labeling roads, streams, etc., and 'spotting' each structure listed by them on this sketch map.

Table 1
Estimated Missed Housing Units per 100 Enumerated Units, 1970 Census of Housing and Population

	Total Missed Units	Missed Units In Enumerated Addresses	Missed Units in Missed Addresses
Total U.S.	2.5	0.5	2.0
Rural	4.8	0.2	4.6
Urban	1.7	0.6	1.1
In Urbanized Area	1.3	0.5	0.8
Outside Urbanized Area	3.1	0.8	2.3
Places by Size			
1,000,000 and over	1.1	0.8	0.3
500,000-999,999	0.1	N.A.	0.1
250,000-499,999	0.9	0.4	0.5
100,000-249,999	2.3	1.8	0.6
50,000-99,999	1.7	0.4	1.3
25,000-49,999	2.0	0.6	1.4
10,000-24,999	1.5	0.7	0.8
2,500-9,999	3.2	0.5	2.6

Note: Above are field enumeration coverage rates only (before corrections made in processing). They are taken from The Coverage of Housing in the 1970 Census, U.S. Bureau of the Census, Census of Population and Housing; 1970, Evaluation and Research Program PHC(E)-5. Sampling errors and descriptions of methodology and limitations of the 1970 housing coverage studies appear in the report PHC(E)-5.

In many cases, matching was impossible because names of 1970 household heads were missing or incorrect on the Successor Check listings and map spottings (particularly for the 1970 Census listings) were absent, inaccurate or illegible.[9] It was necessary, therefore, to send over a third of the SC forms back to the field for 'reconciliation'. In the reconciliation, the interviewer was told to try to obtain more definitive information (primarily names of all possibilities as 1970 household heads for the unit) to determine whether the structure was or was not listed in the 1970 Census Address Register and to continue the SC listings until a total of four successor structures which appeared on the 1970 Census Address Registers had been listed (or until certain cut-offs, established for the original SC listing, had been reached). To avoid having to send back for a second or third 'reconciliation' cases not resolved by the first 'reconciliation', the

reconcilers were given copies of the 1970 Address Register Sheets which contained the sample unit and the structures near it. They were also told to get interviews for the housing units they determined to be missed if there were one or two such units. Interviews were not taken for cases with 3 or more unmatched units because subsequent matching in the office usually indicated that such cases were matched by housing units listed on Census Address Register sheets not supplied to the reconciler.

Discussion

As already noted, a coverage evaluation was considered important for the SIE from both the methodological and substantive standpoints. From the methodological standpoint, the SIE introduced some changes over the CPS sampling procedure and it was, therefore, desirable to check whether the coverage resulting from these changes was satisfactory. From the substantive standpoint, the undercoverage could have an important impact on the count of children in poverty, and its distribution among states (and between urban and rural areas) because of the greater missed rates usually found for low income families. It was, in fact, possible that the content and within household coverage checks carried through on the Reinterview Sample proper would tend to reduce the counts of poverty families and of children in such families. That is, more family incomes will tend to be adjusted upward than downward due to (a) the reporting of previously omitted income sources and (b) due to adding omitted income recipients. It is true that the within housing unit coverage check would tend to increase family size but a large component of within household undercoverage is the omission of adult male wage earners.[10] As opposed to this, children ages 5 to 14 tend to be exceptionally well-enumerated among blacks and, probably, among most other groups. However, household reinterviews are usually unsuccessful in detecting missed adult males in enumerated low income families and there is no assurance that the reinterviews with enumerated households in the SIE sample will adequately measure the effects of content errors and within household coverage error.

In contrast with the upward bias of the estimates of the number of children in poverty families due to errors in income reporting and within household coverage, we would expect a downward bias due to the omission of housing units and, because of the greater omission rates for lower income households, we would also expect downward bias in the estimates of the proportion of all children who are in poverty families.

With respect to the estimated coverage of housing units in enumerated structures, the SIE compares favorably with the Census (and, probably, with the CPS). The estimated rate of missed housing units (per 100 enumerated housing units) in enumerated structures is 0.5%, which is the same as the coverage rate of missed housing in enumerated structures estimated for the U.S. as a whole over the 1970 Census. This may, in fact, represent an improvement of the SIE in the 1970 Census coverage within enumerated structures, since the Within Structure Listing coverage estimate should include some housing units "converted" to residential use after the 1970 Census as well as housing units actually missed by the 1970 Census. It is likely that any difference of this type is due more to sampling and matching error than to improved SIE marksmanship.

The picture for the change from an area to a list sample in rural areas and small towns, is less encouraging. For these areas, preliminary SC coverage check estimates are of the order of 7 to 13 missed housing units in missed structures per 100 enumerated housing units (in enumerated structures).[11] This is greater than the 4.6 missed units per 100 enumerated units in missed structures reported for rural areas in the 1970 Census and the rate of 2.6 missed housing units per 100 reported for urban places of population size 2500 to 9999. There is an excellent chance that the SC missed rate represents a real difference in coverage between a "segment" sample and an address sample (of individual housing units built before 1970) due to 'conversions' and particularly 'conversions' of vacant structures which were considered to be nonresidential or "unfit for human habitation" in 1970 because at the time they were vacant but which were classified as housing units because they were occupied for residential use at the time of the SIE.

If, as is likely, a missed housing rate of 13 per 100 (or even of 7 per 100) is considered unsatisfactory for Bureau of the Census surveys and this makes a straight list sample of (enumerated) addresses infeasible, we may want to consider a successor sample as an alternative to a regular area sample, provided we can solve the cost problems associated with making additional visits to 'reconcile' matching problems and to 'complete the string' of units listed in the original sampling frame. That is, the very marked increase in recent years in the costs of delineating satisfactory area segments for sampling purposes may more than offset the 'successor sample' costs of doing a moderate amount of revisits for 'reconciliation' and 'completing the string' of enumerated structures.

It should be noted that the successor check used in the SIE represents a modification of the 'half-open interval' approach used for some previous coverage checks. In the 'half-open interval' approach, units (if any) from the starting point through the next previously listed unit are in the sample. This was modified for SIE to extend the sample 'segment' through the next four previously listed units since, while extending the listing means increased cost, it also means a more than proportionate reduction in variance. The Census Bureau is planning to analyze the data from the SIE successor check and other successor checks done subsequently, to try to estimate the optimum cluster size (from the cost-variance standpoint).

While the successor checks used to date have been used for checking on the coverage of a housing unit listing, the technique can, of course, be

used for updating an old listing. The procedure could be used for the purpose of list updating without matching to the old list by determining those structures which should have been on the old list. This involves carrying forward to the new listing the undercoverage of the old one. In theory, the procedure is, in other respects, no more biased than the listing of an area segment, since the errors made by successor listers in following the route and defining old and new construction, correspond to the errors made by area listers in defining the area boundaries and covering all the units inside those boundaries and none outside of it.

Footnotes

1/ Problems of Nonsampling Error in the Survey of Income and Education: Content Analysis by Robert E. Fay III.

2/ These conversions are not included in the 'New Construction' strata.

3/ This happened either because of inadequate distinction in the Address Register between the housing units at the address or because of conversions or changes in the housing unit identification system.

4/ Many of these cases are merely failure of the housing unit designations to correspond--e.g., one listing shows 1st floor right, 1st floor left, 2nd floor right, 2nd floor left and the other shows apartments 1, 2, 3, 4.

5/ Where the original SIE sample unit was selected for a coverage check interview, the "net coverage error" was defined as zero and no coverage error interview was taken.

6/ Where the sample unit is in a multi-unit structure, one must also allow for the probability that the structure (or structure 'chunk') will be in the sample. This probability is, of course, proportional to the number of units listed for the structure (or structure 'chunk') in the 1970 Census Address Register.

7/ U.S. Bureau of the Census, Census of Population and Housing: 1970, Evaluation and Research Program PHC(E)-5, The Coverage of Housing in the 1970 Census, U.S. Government Printing Office, Washington, D.C., 1973

8/ As might be anticipated, rural areas and the smaller urban places show low rates for housing units missed at (multi-unit) enumerated addresses.

9/ Many of the Census E.D. maps were of such small scale that it was impossible to distinguish between the locations of individual houses in a row of 5 to 10 successive structures.

10/ This shows up clearly in the much higher undercounts in most U.S. censuses and surveys for black males than for black females in the age range 20 to 54.

11/ The range reflects the serious difficulties (and the resultant uncertainties) encountered in trying to match housing units in areas where information on address or location is missing, vague or erroneous. The figures cited are subject to sampling error.