INNOVATIONS IN THE SAMPLE DESIGN FOR THE SURVEY OF INCOME AND EDUCATION

Donald M. Luery and Gary M. Shapiro, U.S. Bureau of the Census

I. INTRODUCTION

This paper discusses in detail some of the more interesting aspects of the sample design of the Survey of Income and Education (SIE) and should be of prime interest to people engaged in designing complex surveys. Some other aspects of the sample design for this survey are covered in detail in Boisen [1] and are briefly discussed in this memorandum.

Only 6-9 months' time was available to decide on and execute the sample design for this survey. Optimality criteria were generally applied in determining the sample design, but the application was generally imperfect.

The SIE was designed to meet three major objectives. Title I of the Elementary and Secondary Education Act of 1965 [10] provided for the annual distribution of $2,000,000,000 to local school districts, with the intent that school districts servicing low income areas should receive relatively more money than school districts servicing high income areas. One provision of the Educational Amendments of 1974 [11] to this Act states that the Secretary of Commerce shall "expand the current population survey (or make such other survey) in order to furnish current data for each State with respect to the total number of school age children in each State to be counted for purposes of Section 103(c)(1)(A) of Title I of the Elementary and Secondary Act of 1965." Thus, the prime objective for the SIE was its use in conjunction with the existent Current Population Survey[1] (CPS) to produce estimates of children, age 5-17, in poverty families with coefficients of variation of 10 percent or better by State.

Another section of the same law dealt with questions of bilingual education and required the Office of Education in the Department of Health, Education and Welfare (HEW) to issue a report to Congress including among other things, "...a national assessment of the educational needs of children and other persons with limited English-speaking ability ..." (PL93-380). This leads to a secondary purpose for SIE of providing estimates of persons with limited English-speaking ability by State. The questions relating to language ability were to be asked only on the SIE questionnaire, not on the CPS questionnaire, and hence, the language ability tabulations were to be based only on the SIE.

The tertiary objective of the SIE was to provide cross-tabulations involving poverty and other items from SIE by itself that were of interest to analysts in the Department of Health, Education and Welfare. The reason that CPS was not to be used for these tabulations was that the SIE questionnaire contained additional questions on food stamp recipiency, housing costs for homeowners and renters, estimated cash receipts, education, disability, and health insurance coverage.

Initially, SIE was intended to have a designated sample size of 200,000 housing units. The methods used in deciding how this was to be allocated by State, consistent with the three objectives discussed above, are discussed in Section II. For budgetary reasons, it was decided after the basic sample was selected that a sample reduction to about 190,000 designated units, resulting in about 151,000 interviewed households, was needed. This required reallocation of sample by State is discussed briefly in Section II.

The SIE was designed completely independently of the CPS on a State-by-State basis, except that the primary sampling unit (PSU) definitions were the same. In most States, primary sampling units consisting of SMSA's or groups of counties and independent cities, were divided into strata according to estimates based on 1970 census data of the proportion of persons who were children age 5-17, living in poverty families. PSU's in a State that were large enough to provide at least 80 sample housing units formed a stratum by themselves and came into sample with certainty. In nine States (Conn., Del., D.C., Hawaii, Md., Mass., N.H., R.I., and Vt.) every PSU was selected with certainty. In the remaining States, from one to ten non-self-representing strata with three or more PSU's were formed in each State. Two sample PSU's were selected with replacement from each stratum using the Durbin-Sampford rejective method. See Durbin [5] and Sampford [7].

The major frame for sampling housing units from a selected PSU was the list of units enumerated in the 20 percent sample of the 1970 census. The 20 percent sample was used instead of the full census file because of the information on income and poverty available from it. Two methods of selection were employed in the selection from the census file. For the first method, some enumeration districts (ED's) were selected and a sample of approximately three housing units was selected from each ED. (An ED was the assignment given to a single interviewer in the 1970 census. On the average, an ED contains approximately 350 housing units). For the second method, a direct selection of housing units was taken without the intervening step of selecting ED's. These two methods of selection are more fully described in section IV of this paper. In order to attempt full coverage of housing units, a systematic sample from four additional frames was selected: (1) special places, (2) units built since the 1970 census in jurisdictions that issue permits, (3) units built since the 1970 census in jurisdictions that do not issue permits, and (4) mobile homes in parks established since the 1970 census.

Section III of this paper discusses the methods used to decide that noncompact clusters of three housing units should be used for most States.

Section IV discusses why the Durbin method was used for most stages of selection, how it was applied, the difficulties caused by the required reduction in sample size, and some advantages and disadvantages of the Durbin method.

II. ALLOCATION OF SAMPLE BY STATE

Sample was allocated to each State in accordance with the three primary objectives of the survey as stated above and the amount of money available for the survey. Most of the credit for the allocation scheme which is described should go to Mr. Wray Smith in the Office of the Secretary, Department of Health, Education, and Welfare. The

authors assume full responsibility, however, for any errors in this paper and any problems of logic with the allocation scheme.

The sample was not allocated in one stage but rather in three stages, one stage for each primary objective. The vast majority of the sample was allocated to satisfy the first objective of producing estimates of children, age 5-17, in poverty families. However, this was done in the first stage, so that the allocation decisions for the other two objectives could take advantage of the large sample intended to satisfy the first objective. Had sample been allocated in one stage or in a different order to meet the three objectives, there would have been substantial differences in sample size for some States.

We began with the sample present in the CPS including the supplementation to CPS begun in July 1975. (See Dippo [4] for details on this supplementation.) The sample totals are given by State in column (2) of table 1. We determined the additional sample needed for each State to achieve an expected 9.6 percent coefficient of variation on the estimated children, aged 5-17, in poverty families in the State. The choice of 9.6 percent was somewhat arbitrary. The criteria had to be a coefficient of variation less than or equal to 10.0 percent; 9.6 percent was an affordable criteria and brought a little bit of safety for achieving a true 10.0 percent coefficient of variation in each State. A number of assumptions were needed to determine the sample sizes. Perhaps the most important was an estimate of the number of children in poverty families. Rather standard methodology was used, however, so no description will be given here. Details are given in appendix A of Boisen [1] and there is some related discussion in section III of this paper. The supplemental sample sizes to meet this objective are given by State in column (3) of table 1.

Next we allocated about 36,000 sample households to the States to improve the estimates of persons with difficulty speaking English. These estimates were to be made from the SIE sample only. The 36,000 figure corresponded roughly with the amount of money being contributed to the total survey effort by the part of HEW interested in these estimates. This additional sample was allocated in order to bring the total allocation closer to optimal allocation, according to the standard optimum allocation formula, for a national estimate of persons with difficulty speaking English. The second objective is, of course, concerned with State, not national estimates. However, there was no requirement for equal reliability for each State and, in fact, it was felt that States with a relatively serious problems of persons with difficulty speaking English needed greater reliability in their estimates. Optimally allocating a sample for a national estimate is one way of achieving this. At the same time, it was desired that all States have a reasonably large sample size for the planned analysis and 2,000 was selected as a minimal supplementary sample size per State for this purpose.

Finally, we allocated sample households to the States to improve estimates of children in poverty based on the SIE sample only, without benefit of the CPS sample. The criteria was a 9.9 percent CV on State estimates. The third objective does not relate specifically to total children in poverty and the choice of 9.9 percent CV is completely arbitrary other than it being consistent with the total sample size that could be afforded. It was felt, however, that this allocation would well serve the third objective. Sample sizes are given in column (5) of table 1.

All of the above relates to the original allocation before the budget-imposed reduction. The reallocation necessitated by the reduction was accomplished through similar procedures. Instead of a 9.6 percent CV criteria for the first allocation, a 9.8 percent CV criteria was used; this reduced the sample allocated in this stage from 157,000 to 148,000. The procedure and number of sample cases for the second stage of allocation was completely unchanged. In the third stage of allocation the criteria was 10.4 percent CV instead of 9.9 percent CV; this reduced the sample allocated in this stage from 12,000 to 6,500. The final supplementary sample sizes after reduction are given in column (10) of table 1.

Note that all sample sizes given are originally intended expected sample sizes. The figures were used to determine sampling rates. Application of these sampling rates did not yield the exact figures given in column (10).

III. DETERMINATION OF NONCOMPACT CLUSTERS OF THREE HOUSING UNITS

We started with the assumption that we would generally select a sample of enumeration districts (ED's) from the sample of PSU's and that only a single cluster of housing units (or a single special place hit) would usually be selected from each ED. In order to objectively determine optimal cluster size and to determine whether clusters of housing units should be compact or dispersed throughout an ED (noncompact), several cost figures and intraclass correlations were necessary.

Comparisons for different compact and noncompact cluster sizes were based on estimated design effects; that is, estimates of the increase in variance because cluster sampling of households instead of a simple random sample of persons was used. The characteristic of interest was school-aged children in poverty families. In all the calculations made, we assumed a simple random sample of ED's from the State and, for noncompact clusters, a simple random sample of housing units from ED's. In fact, of course, ED's were not generally selected directly from a State and a systematic sample of housing units was selected for noncompact clusters.

The formula used for the design effect for compact clusters, which measures the increase in variance expected from selecting compact clusters of households as compared to selecting a simple random sample of persons, was[2]

$$(\hat{V}_L^2/\hat{V}_L^2)\ (\hat{V}_K^2/\hat{V}_K^2)\left[1+\delta_{\bar{P}}(\bar{P}-1)\right] \qquad (1)$$

The formula used for the design effect for noncompact clusters of housing units versus a simple random sample of persons was[3]

$$(\hat{V}_L^2/\hat{V}_L^2)\ (\hat{V}_K^2/\hat{V}_K^2)\left[1+\delta_{\bar{\bar{K}}}(\bar{\bar{K}}-1)\right]\left[1+\delta_{\bar{\bar{N}}}(\bar{\bar{N}}-1)\right] \qquad (2)$$

TABLE 1.—Sample Sizes (Housing Units) by State and by Stages of Allocation, and Coefficients of Variation for Important Estimates

STATE	CPS Sample Size (including Supplementation)	Supplement for Children in Poverty Estimate	Supplement for Bilingual Estimates	Supplement for Ests. Based on Sample Excluding CPS	Total Supplementary Sample Size (3)+(4)+(5)	CV's for Children in Poverty Est. Based on Complete Sample	CV's for Children in Poverty Based on Sample Excluding CPS	CV's for Persons with Difficulty Speaking English	Total Supplementary Sample Size After Reduction
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
TOTAL	68790	157253	35969	11906	205127				190243
Maine	900	2512	0	374	2886	0.091	0.099	0.088	2747
New Hampshire	740	5285	0	132	5417	0.095	0.099	0.057	5252
Vermont	670	3413	0	161	3574	0.094	0.099	0.082	3370
Massachusetts	1450	4251	233	420	4903	0.091	0.099	0.055	4458
Rhode Island	580	4212	0	99	4311	0.095	0.099	0.054	4032
Connecticut	890	5479	0	229	5707	0.094	0.099	0.046	5175
New York	4680	1057	4374	0	5431	0.068	0.086	0.043	5221
New Jersey	1850	3779	2042	0	5821	0.081	0.089	0.045	5666
Pennsylvania	3070	2920	2778	0	5698	0.077	0.091	0.065	5464
Ohio	2750	3307	2446	0	5754	0.079	0.090	0.075	5502
Indiana	1550	5201	0	744	5946	0.091	0.099	0.084	4794
Illinois	2770	2704	2963	0	5667	0.075	0.088	0.057	5465
Michigan	2370	3330	2427	0	5757	0.079	0.090	0.067	5514
Wisconsin	1070	4783	0	404	5187	0.093	0.099	0.057	3966
Minnesota	1250	4304	0	473	4778	0.092	0.099	0.065	4084
Iowa	950	4720	0	261	4982	0.094	0.099	0.081	4535
Missouri	1540	2596	71	652	3319	0.087	0.099	0.112	3000
North Dakota	980	3198	0	425	3624	0.091	0.099	0.066	4143 *
South Dakota	1120	1806	75	493	2373	0.087	0.099	0.096	2877 *
Nebraska	800	3651	0	238	3889	0.093	0.099	0.082	3603
Kansas	880	3991	0	279	4270	0.093	0.099	0.091	3979
Delaware	540	3555	0	183	3737	0.094	0.099	0.093	2910
Maryland	980	3663	0	461	4124	0.091	0.099	0.090	3160
District of Columbia	550	2206	0	217	2423	0.092	0.099	0.100	2171
Virginia	1230	2663	0	565	3229	0.089	0.099	0.131	2569
West Virginia	840	2445	0	339	2784	0.091	0.099	0.161	2212
North Carolina	1310	1785	314	274	2373	0.086	0.099	0.180	2114
South Carolina	830	1522	588	0	2110	0.085	0.096	0.200	2000
Georgia	1330	1377	654	41	2071	0.084	0.099	0.177	2000
Florida	2320	2179	961	275	3415	0.083	0.099	0.070	3310
Kentucky	910	1965	0	314	2279	0.090	0.099	0.178	2000
Tennessee	1010	1838	152	335	2325	0.088	0.099	0.182	2175
Alabama	1030	1320	846	0	2166	0.080	0.091	0.202	2199 *
Mississippi	810	766	1187	47	2000	0.068	0.077	0.203	2000
Arkansas	870	1603	354	44	2000	0.088	0.098	0.196	2000
Louisiana	1130	1033	1624	0	2657	0.068	0.076	0.087	2165
Oklahoma	960	2177	0	410	2586	0.090	0.099	0.137	2476
Texas	3270	588	4775	0	5363	0.054	0.062	0.039	5182
Montana	1010	3797	0	327	4124	0.093	0.099	0.083	3823
Idaho	900	6054	0	143	6198	0.095	0.099	0.072	5807
Wyoming	720	4959	0	142	5102	0.095	0.099	0.066	4253
Colorado	970	3352	0	385	3737	0.092	0.099	0.059	4130 *
New Mexico	910	1083	2063	0	3146	0.063	0.068	0.034	2580
Arizona	800	2649	0	334	2983	0.091	0.099	0.052	2657
Utah	900	4612	0	362	4974	0.093	0.099	0.076	5057 *
Nevada	650	5323	0	89	5418	0.095	0.099	0.058	4911
Washington	1000	4646	0	328	4974	0.093	0.099	0.075	4339
Oregon	880	4480	0	231	4711	0.094	0.099	0.079	4896 *
California	5690	278	5041	0	5319	0.065	0.086	0.043	5117
Alaska	1030	2395	0	515	2909	0.089	0.099	0.082	3780 *
Hawaii	550	4434	0	162	4596	0.094	0.099	0.050	3401

NOTE: The first 9 columns of this table appeared in Boisen [2]; column (10) appeared in Smith [8].

* The sample sizes are higher for these States after the reduction allocation than from the original allocation because more accurate data on between PSU variances was available and the variances for these States were higher than previously speculated.

The notation and meaning of these terms is as follows:

V_K^2 is the population relvariance between persons for the proportion of poverty children.

$\hat{V}_K^2$ is like V_K^2 except it includes the effect of the number of persons per listing unit.

$(\hat{V}_K^2/V_K^2)$ is the increase in variance due to variation in the number of persons per listing unit. Note that the listing unit is taken to be a compact cluster of housing units for Formula (1) and a single housing unit for Formula (2). Thus, $(\hat{V}_K^2/V_K^2)$ is not precisely the same quantity in (1) as in (2).

V_L^2 is the population relvariance between housing units for the number of poverty children per household.

$\hat{V}_L^2$ is like V_L^2 except it includes the effect of the variation of the number of housing units per ED.

$(\hat{V}_L^2/V_L^2)$ is the increase in variance due to variation in the number of housing units per ED.

$\bar{\bar{K}}$ is the average number of persons per housing units.

$\bar{\bar{N}}$ is the average number of housing units per cluster.

$\bar{\bar{P}}$ is the average number of persons per cluster. $\bar{\bar{P}} = \bar{\bar{K}}\,\bar{\bar{N}}$.

$\delta_{\bar{\bar{P}}}$ is the intraclass correlation between persons within listing units. For noncompact clusters, $\bar{\bar{P}}=\bar{\bar{K}}$.

$1+\delta_{\bar{\bar{P}}}(\bar{\bar{P}}-1)$ is the increase in variance due to sampling listing units instead of persons, assuming no variation in number of housing units per cluster.

$\delta_{\bar{\bar{N}}}$ is the intraclass correlation between housing units within an ED.

$1+\delta_{\bar{\bar{N}}}(\bar{\bar{N}}-1)$ is the increase in variance due to sampling a cluster of $\bar{\bar{N}}$ housing units instead of a single housing unit.

For calculating $\delta_{\bar{\bar{K}}}$ (for a compact cluster of one housing unit), we obtained a special tabulation from the 1970 decennial census giving, for each family size, the distribution of families with zero children in poverty, one child in poverty, two children in poverty, etc. From this we were able to calculate directly a within family-size group relvariance W^2 and a between family-size group relvariance B^2. From these relvariance estimates, we calculated:

$$\delta_{\bar{\bar{K}}} = \frac{\frac{N-1}{N}B^2 - \frac{W^2}{\bar{\bar{K}}}}{\frac{N-1}{N}B^2 + \frac{(\bar{\bar{K}}-1)W^2}{\bar{\bar{K}}}} \qquad (3)$$

where N is the number of housing units in the U.S., hence $\frac{N-1}{N} \doteq 1$.

We needed $\delta_{\bar{\bar{P}}}$ for values of $\bar{\bar{P}}$ other than $\bar{\bar{P}}=\bar{\bar{K}}$ as well. From census data at the ED level, we were able to estimate $\delta_{\bar{\bar{P}}=900}$, where 900 was the average number of persons in an ED, by a procedure described below. We then used these two calculated $\delta_{\bar{\bar{P}}}$'s to fit the curve $\delta_{\bar{\bar{P}}} = a\bar{\bar{P}}^b$ (p. 307 of Hansen, Hurwitz & Madow [6]). Having estimates of $\delta_{\bar{\bar{P}}}$ for only two values of $\bar{\bar{P}}$ is, of course, not very satisfactory, but we could do no better because of time restrictions. The computed values are as follows:

$$\delta_{\bar{\bar{P}}=3.1} = .55$$

$$\delta_{\bar{\bar{P}}=900} = .06$$

$$a \doteq .85$$

$$b \doteq -.4$$

With considerable effort, we were able to estimate $\delta_{\bar{\bar{P}}}$, the intraclass correlation between persons within ED's, from census data at the ED level. We needed some special tabulations from the 1970 decennial census, but time and money constraints prohibited running the entire census file, so calculations were made initially only for Wisconsin. (Calculations were subsequently made for Georgia and generally confirmed the earlier results.) There probably are some significant differences between the intraclass correlations for some States and those for Wisconsin and there may also be nontrivial changes in the intraclass correlations from 1970 to 1976, though these latter differences would not necessarily affect the optimum noncompact cluster size.

From the 20-percent census[4] data at the ED level, we computed:

$$\delta_{\bar{\bar{P}}=900} = \frac{S_1^2 - \bar{\bar{K}}\,S_2^2}{S_1^2 + \bar{\bar{K}}(\bar{\bar{K}}-1)\,S_2^2} \qquad (4)$$

where $S_1^2 = \frac{1}{M_s-1}\left\{\sum_{i=1}^{M_s} X_i^2 - \frac{1}{M_s}\left(\sum_{i=1}^{M_s} X_i\right)^2\right\}$

$$S_2^2 = \frac{1}{K_s}\sum_{i=1}^{M_s}\frac{X_i(K_i-X_i)}{K_i-1}$$

K_s is the State population,

M_s is the number of ED's in the State,

$\bar{\bar{K}} = \frac{K_s}{M_s}$, the average number of persons per ED,

K_i is the population of the i^{th} ED,

and X_i is the number of poverty children in the i^{th} ED.

The final quantities needed for computing the design effect for compact clusters (Formula (1)) are $(\hat{V}_K^2/V_K^2)$ and $(\hat{V}_L^2/V_L^2)$. Both these ratios are are assumed as constants in the calculations. In fact, however, they are somewhat a function of the cluster size. For $(\hat{V}_L^2/V_L^2)$ fewer clusters mean a larger number of ED's would turn out to be self-representing and would not contribute to (V_L^2/V_L^2). (In the extreme where all ED's are self-representing, $(\hat{V}_L^2/V_L^2)=1.0$, otherwise it is greater than 1. Since ED's are selected with a probability based on their size, (V_L^2/V_L^2) is expected to be close to 1.0.) For $(\hat{V}_K^2/V_K^2)$, the relative variation in number of persons per cluster is likely to decrease with the size of the cluster since large households will be combined with small households in larger clusters. Also, for characteristics of a small proportion of the total population, this quantity is not appreciably affected by the size of the cluster. (V_L^2/V_L^2) is assumed as a constant and thus left out of the

computations entirely. $(\hat{V}_K^2/V_K^2)$ could have been treated similarly, but instead was speculated as a constant 1.3 and carried through in the computations. Design effects for different compact cluster sizes are given in table 2.

TABLE 2. DESIGN EFFECTS FOR DIFFERENT CLUSTER SIZES

Cluster Size and Type	Design Effect
1 Housing Unit	2.8
COMPACT:	
2 Housing Units	4.0
3 Housing Units	5.0
NONCOMPACT:	
2 Housing Units	3.0
3 Housing Units	3.2
4 Housing Units	3.5
5 Housing Units	3.7
6 Housing Units	3.9

For noncompact clusters, $\delta_{\bar{N}}$ was estimated by[5]

$$\delta_{\bar{N}} = \frac{[1+\delta_{\bar{P}=900}(\bar{K}-1)] - [1+\delta_{\bar{P}=3.1}(\bar{K}-1)]}{(\bar{N}-1)\ [1+\delta_{\bar{P}=3.1}(\bar{K}-1)]} \qquad (5)$$

where

$\bar{N}$ is the average number of housing units per ED. Other terms were defined earlier.

The calculations yield $\delta_{\bar{N}} = .08$. Using this value in Formula (2) resulted in the design effects for noncompact clusters also shown in table 2.

We decided that any cost advantages for compact versus noncompact clusters were not sufficient to make up for the sizable design effect differences as shown in table 2, and then proceeded to determine the optimal cluster size for noncompact clusters. Table 3 compares the variable costs for additional interviews and interviewers to be incurred for alternative cluster sizes for a particular portion of the country. (Total survey costs run much higher.) The portion represented is the full States of Maryland and Massachusetts, and Milwaukee, Dane, and Brown counties in Wisconsin. The main reason for the choice of these particular areas is that data for direct field costs happened to be readily available for them.

Calculations were made for each cluster size in each of the five areas separately, and then summed to produce table 3. Consider Maryland, for example. For a given cluster size, the appropriate design effect was used to determine the number of sample units needed to achieve a CV of 10 percent on the estimated number of children in poverty families. The 1970 census figure on children in poverty families was used for the level of the estimate. The number of interviewers required for such a sample size was then estimated, which in turn determined the cost of training and recruiting. It was assumed, based on prior survey experience, that an additional interviewer is required for each increase of 100 units in sample. The training and recruiting cost was $350 per interviewer. Sampling costs were mostly a function of the number of ED's in sample. Field costs represent the direct interviewing costs. The table shows that the cost was minimized for clusters of three and thus this is what we used in the actual survey.

For the three counties in Wisconsin, we determined the number of sample units required for the State as a whole and then allocated this down for the three counties of interests. Cost figures were then developed separately for each county in the same manner as for the two States.

We made another set of computations that tended to confirm the estimated $\delta_{\bar{P}=3.1}=.55$ and $(V_K^2/V_K^2)=1.3$. These computations were also based on the special tabulation from the 1970 census giving the distribution of families with 0 children in poverty, 1 child in poverty, etc. From this distribution, we calculated an estimate of the population relvariance between households of the number of children aged 5-17 in poverty per household; this relvariance is V_L^2 defined previously. We also determined the population relvariance between persons for the proportion of total persons that are poverty children aged 5-17. This is the same as V_K^2 defined previously where $V_K^2 = (1-P)/P$, where P is the proportion of poverty children.

To compare V_L^2 with V_K^2, which can be considered as the relvariances for a simple random sample of one household and one person respectively, V_L^2 needs to be adjusted by the average number of persons per household. Therefore the design effect for a simple random sample of n_H households versus a simple random sample of $\bar{K}n_H$ persons, with $\bar{K}=3.1$, is given by, $\bar{K}\ V_L^2/V_K^2$. This design effect was approximately 2.8. Although this procedure involves less computations; it does not produce a value for the intraclass correlation between persons within a household that was used in other aspects of the sample design.

IV. DURBIN-SAMPFORD SAMPLE SELECTION

As stated previously, the prime objective of SIE was to produce estimates of children, age 5-17, in poverty families with coefficients of variation no worse than 10 percent for each State. It was felt that reliable estimates of variance were desirable in order to verify that we had met the specified reliability requirements and that good estimates of variance be available for the analysis of the data resulting from the survey. The sample selection can be divided into two stages: First, the selection of PSU's and second the selection of housing units from PSU's. The discussion will be divided between these two stages.

Selection of Primary Sampling Units. Generally when there is sufficient auxiliary information (usually from the most recent census) to enable us to stratify the PSU's, it is assumed that only one PSU needs to be selected from each stratum. Under this assumption of sufficient auxiliary information, the between PSU component of variance is felt to be smaller when forming small strata from which one PSU is selected than when forming larger strata from which more than one PSU is selected. Also, the method of estimating variance when one PSU per stratum has been selected, collapsed strata, produces biased estimates of variance. This bias may be large enough so that the expected value of the variance estimate for collapsed strata is greater than the unbiased estimate of variance for a design

specifying two PSU's per stratum and where the strata are larger than the one PSU per stratum design.

Two considerations entered into our decision whether to select one or two PSU's per stratum. First, the between PSU component of variance needed to be small enough so that the coefficient of variation for estimating poverty children remained less than 10 percent. Preliminary estimates indicated that the between PSU variance for two PSU's per stratum would range between 0 and 30 percent of the total variance (with three-quarters of the States below 10 percent) but that the coefficient of variation would remain below 10 percent for all States but one. Hence, using the Durbin [5] technique to select two PSU's per stratum would not raise the variances above the stated requirements. Second, we found that since the between PSU variance was a minor component of the total variance, the risk was small that we would estimate a coefficient of variation greater than 10 percent when it actually was less. If the risk of estimating a coefficient of variation greater than 10 percent had been high, we might have selected the procedure with the lowest variance since one PSU per stratum or two PSU's per stratum would have both been relatively unattractive. From this preliminary analysis, we concluded that there was no strong reason to prefer one PSU per stratum over two PSU's per stratum, and, as a result, we selected the procedure which would provide unbiased estimates of variance.

Table 4 illustrates an interesting relationship between the between PSU variance for Durbin technique, for one PSU per stratum, and for the collapsed stratum estimate of variance. These calculations were performed subsequent to the decision to use Durbin technique to select PSU's and had no bearing on that decision. Table 4 shows estimates from census data of the between PSU variance based on stratification by 1970 poverty children for six States. Part a. shows the between PSU variance for estimates of 1970 poverty children, that is, an item with perfect correlation with the auxiliary information. The expected relationship is evident in this part of the table, that one PSU per stratum is superior to two PSU's per stratum in terms of true variance but that the expected value of the collapsed stratum estimate of variance exceeds the true variance for two PSU's per stratum. Parts b. and c. of the table show the between PSU variance for 1970 poverty families, an item highly correlated with the auxiliary information, and for 1960 poverty families, an item that shows the effect of displacement in time. (Note that the number of 1960 poverty children is not available for comparison.) Part c. no longer shows a clear advantage for one PSU per stratum over two PSU's per stratum from larger strata. The collapsed stratum estimate of variance still provides an overestimate of the true variance though its relative overestimate is less. Taking into consideration the fact that the auxiliary information will not be perfectly correlated with the key items of a survey, Table 4 indicates that two PSU's per stratum may be a good compromise between the reduction in variance due to stratification and obtaining an unbiased estimate of variance.

Selection Within PSU's. The most common method used at the Bureau of the Census to select a sample of units from within a PSU is to take a sorted systematic sample. This method has the obvious advantages of being easy to implement and of resulting in a relatively low true variance for items correlated with the sort variables, but has the disadvantage that, since the systematic sample is in effect a sample of one cluster per PSU, no unbiased estimate of the variance exists. The methods to estimate the variance will tend to overestimate it when the sort variables are effective in reducing the variance.

Along with an estimate of variance due to the selection of PSU's as described above, we required an estimate of the within sampling variance for the self-representing sample PSU's and, since we were using the Durbin technique to select the non-self-representing PSU's, we required an estimate of the within sampling variance for each of the non-self-representing PSU's. We felt that the application of the Durbin technique to the within PSU sample selection would lead to little, if any, increase in variance over a systematic sample[6] since (1) the strata would be formed in the same sort as if a systematic sample were to be used, and (2) the strata would be numerous and small so that each stratum would be fairly homogeneous. Also, selecting a sample or estimating variances using the Durbin technique is not much more difficult on the computer than other procedures. Thus, we decided to select our sample from the 1970 census using the Durbin technique so that we would be able to estimate the gain in variance due to the sorting and stratification of the sample.

Approximately 85 to 90 percent of our sample in a State came from the 1970 census file and the Durbin technique was used to select the sample. The remainder of the universe, primarily new construction, required clerical operations for the sample selection. As a result, we chose systematic samples from this part of the universe. Since a systematic sample would provide little gain in variance for this part of the sample, the variance was estimated as if it were based on a simple random sample.

The sample selection within a PSU was briefly sketched in the introduction. As described in Section III, above, we decided to select a sample of ED's from which a noncompact cluster of three housing units would be selected from each one. Many ED's were large enough so that we expected them to enter sample with certainty. As a result, we decided to directly select a sample of housing units from these large ED's using the Durbin technique. The housing units from large ED's were sorted by their poverty level and the number of children under 18, and within these by county and ED. In this sort, the housing units were grouped into strata (called Durbin housing unit strata), and two units per stratum were selected using the Durbin-Sampford rejective method.

The remaining smaller ED's were sorted by five size categories and, within each size category, by the percent of persons in poverty such that the first size category was sorted from highest to lowest poverty, the second size category from lowest to highest poverty, etc. In this sort, the ED's were grouped into 12 or more strata

(called Durbin ED strata), and two ED's per stratum were selected using the Durbin-Sampford rejective method where the measure of size was proportional to the number of housing units plus the number of persons in special places divided by three. From a selected ED either a special place or a cluster of three housing units was selected. In either case a systematic sample was taken. It was thought that a substantial improvement in variance could be achieved if the housing units within an ED were sorted by poverty level and number of children less than 18, and a systematic sample of three housing units was taken.

Departures from an Unbiased Estimate of Variance. Approximate methods need to be applied to estimate the variance from those frames from which systematic samples were selected. Thus, though we attempted to select the sample so that we would have an unbiased estimate of variance, we did not achieve this fully. Furthermore, we departed from an unbiased estimate of variance in two additional ways.

First, an estimate of variance for a non-self-representing stratum k is:

$C_k(\hat{X}_{k1}-\hat{X}_{k2})^2 + (1-C_k)\ (\hat{\sigma}^2_{k1} + \hat{\sigma}^2_{k2})$ where

$\hat{X}_{k1}$ and $\hat{X}_{k2}$ are sample estimates from the sample PSU's from stratum k,

$\hat{\sigma}^2_{k1}$ and $\hat{\sigma}^2_{k2}$ are estimates of the within sampling variance for the two PSU's, and

C_k is a constant which is a function of the joint probability of selecting the two sample PSU's and of the probabilities of selecting each of the PSU's in stratum k.

Durbin [5] recommends that the coefficient C_k in the variance formula be reduced to one whenever it exceeds one in order to reduce the variance on the variance estimate. C_k exceeds one when the measures of size of the units in a stratum are diverse and one of the larger units is not in the pair of selected units. Since, the measures of sizes of PSU's in a stratum were rather heterogeneous for this survey, the C_k's were reduced to one according to Durbin's recommendation.

Table 5 shows estimates of the relative bias due to reducing C_k to one, the relative variance of the unbiased estimate of variance, and the relative mean square error of the biased estimate of variance for estimating 1970 poverty children for six States. The relative bias is in general small and there is a decrease in the relative mean square error.

Second, the selection of census housing units was from the 1970 census 20-percent sample. By using the Durbin method to select the sample of housing units from large ED's in the first method of selection, an estimate of the within component of variance due to the 20-percent census sample was required. For the State of Wyoming with an overall sampling rate of about 1 in 25, ignoring this within component of variance would produce a substantial underestimate of the variance (approximately 16 percent). Thus, we decided to estimate variances under the assumption that we had selected a stratified simple random sample from all units represented by the 20-percent census sample. This procedure can be defended as follows: The Durbin sampling from the 20-percent census sample is approximately simple random sampling because the measures of size were, in general, nearly equal. If we assume that the 20-percent census sample was a simple random sample, then we can conclude that, overall, we had selected a simple random sample of households.

Reduction in Sample. Section II of this paper discusses how the sample was reallocated due to the budget-imposed reduction. There was no reduction in nine States and a reduction from approximately 2 to 24 percent in the remaining States. Because we had selected our sample of ED's or of housing units using the Durbin method, the reduction in sample was complicated.

The reduction in cost could be achieved in two ways: (1) by eliminating interviews and (2) by reducing the number of ED's an enumerator would have to visit. Thus we could reduce the cost both ways if we deleted every sample housing unit from an ED. Reducing ED's from the Durbin ED selection was no problem. A systematic sample of Durbin ED strata was selected and one of the two sample ED's was randomly deleted with equal probability. It was thought that the increase in variance due to deleting one of two ED's from a stratum was less than the increase in variance from deleting both ED's from half as many strata; no estimates were made of this difference.

For the housing unit selection using the Durbin procedure, it was thought that there could be an excessive increase in variance if all housing units in a sample ED were deleted because of the clustering of the sample housing units. Table 6 shows for States in which a large part of the housing units had been directly selected using the Durbin procedure, the average number of sample households per ED, the approximate percent reduction required, the increase in variance of an ED reduction over a simple housing unit reduction, and the expected CV's after an ED reduction. Four States had an excessive increase in variance from an ED reduction that brought their expected CV's over 10 percent. For each of these four States (Delaware, Wyoming, Nevada, and Hawaii), a systematic sample of Durbin housing unit strata was selected and both housing units from a selected stratum were deleted for the reduction. For the remaining States, the ED's with sample housing units selected using the Durbin procedure were ordered by the number of sample housing units in the ED and a systematic sample of ED's was deleted for the reduction. Note that, because of the underlying random structure of the Durbin sample selection, every pair of ED's from the Durbin ED selection and every pair of housing units from the Durbin housing unit selection retains a positive joint probability of selection in spite of the systematic reduction. Hence, an unbiased estimate of variance (except as noted above) can still be obtained. The derivation of the unbiased estimates of variance has been completed but their presentation would be rather complicated and they will not be included in this paper. Documentation has not been completed.

Advantages and Disadvantages of the Durbin-Sampford Selection Method. The Durbin-Sampford method of sampling selection has been discussed in two contexts in this paper. First, the selection of PSU's, comparing the Durbin procedure with one PSU

per stratum, and second the selection from within sample PSU's, comparing the Durbin procedure with systematic sampling.

The disadvantages of using Durbin procedure, with respect to one PSU per stratum and systematic sampling, are approximately the same. First, it is more difficult to select the sample using the Durbin-Sampford method. This is rather small when it is implemented on the computer but it could be a very difficult task when selecting the sample by hand if there were to be numerous strata. Second, the estimate of variance is more complicated since a constant for each stratum has to be calculated and additional components of variance usually need to be estimated. This increase in the difficulty of estimating variances can often be reduced if the variance estimate can be adapted to replications. Durbin [5] points out a method to estimate the variance which can be easily adapted to the replication method of estimating variances. Third, the true variance from the Durbin procedure may be larger when an item is highly correlated with the auxiliary information that was used for sorting or stratification. Finally, a sample selected using the Durbin procedure is less versatile if a further supplementation or reduction in the sample is required.

The most obvious advantage of the Durbin method is that it provides an unbiased estimate of variance. It appears to be a reasonable balance between providing an unbiased estimate of variance and reducing variance by sorting and stratification of the universe. This would in general be true for any scheme which selects two units per stratum. The advantage of the Durbin procedure over other without replacement schemes is that the Durbin-Sampford rejective method is comparatively easy to implement and that the constants in the variance estimate are directly calculable from terms used in the sample selection procedure. Sampling with replacement is easier to implement but will produce variances larger than Durbin.

A second less obvious advantage for the Durbin method is that the estimate of variance may itself have a lower variance than the estimated variance from the collapsed stratum technique. This can be argued as follows. An estimate of total variance from Durbin selection is:

$$\sum_{k}^{NSR} C_k(\hat{X}_{k1}-\hat{X}_{k2})^2 + \sum_{k}^{NSR} (1-C_k)(\hat{\sigma}^2_{k1}+\hat{\sigma}^2_{k2}) + \hat{\sigma}^2_{SR}, \qquad (1)$$

($\hat{\sigma}^2_{SR}$ is an estimate of the within sampling variance for the self-representing PSU's). The second and third terms, for our survey, have a considerable lower variance than the first terms since they are made up of the sum of squares from numerous strata and the first term is made up of at most 10 sum of squares. Now the expected value of the first term is:

$$\sum_{k}^{NSR} E\, C_k(\hat{X}_{k1}-\hat{X}_{k2})^2 = \sum_{k}^{NSR}\sum_{i>j}\sum(\pi_{ki}\,\pi_{kj}-\pi_{kij})(X_{ki}-X_{kj})^2 + \sum_{k}^{NSR}\sum_{i}\pi^2_{ki}\sigma^2_{ki} \qquad (2)$$

where π_{ki} and π_{kj} are the probabilities of selecting PSU_i and PSU_j from stratum k, π_{kij} is the joint probability of selecting both PSU's i and j, and $X_{ki}=E(\hat{X}_{ki}|\text{the selection of } PSU_i)$. Thus, the first and least accurate term estimates the between PSU component of variance and part of the within component of variance for the non-self-representing stratum. The remainder of the within variance is estimated by the more accurate second and third terms. Similarly, the estimate of total variance from collapsed stratum is $\sum_{k}^{\text{Paired Strata}} (\hat{X}_{k1}-\hat{X}_{k2})^2 + \hat{\sigma}^2_{SR}$ and the expected value of the first term

$$\sum_{k}^{\text{Paired Strata}} E(\hat{X}_{k1}-\hat{X}_{k2})^2 = \sum_{k}^{\text{Paired Strata 2}}\sum_{j=1}^{2}\sum_{i} P_{kji}(X_{kji}-\bar{X}_{kj})^2 + \sum_{k}^{\text{Paired Strata 2}}\sum_{j=1}^{2}\sum_{i} P_{kji}\sigma^2_{kji} + \sum_{k}^{\text{Paired Strata}} (\bar{X}_{k1}-\bar{X}_{k2})^2 \qquad (3)$$

where P_{kji} is the probability of selecting PSU_i from j^{th} strata of the paired stratum denoted by k.

$X_{kji} = E(\hat{X}_{kji}|\text{the selection of } PSU_i)$, and

$\bar{X}_{kj} = \sum_{i} P_{kji}X_{kji}$.

Thus the collapsed stratum term estimates the between PSU component of variance, the within component of variance for NSR strata and the bias from using the collapsed stratum estimator. Again the first term is the least accurate term in the estimate of variance since it would be made up of, at most, 10 sums of squares for SIE. Thus, the least accurate first terms in the estimates of variance, estimate more of the total variance for collapsed stratum than for the Durbin procedure. Thus, the variance estimate may be less accurate for collapsed strata. Research into the variance of our variance estimates will have to be conducted before a conclusive statement can be made that the Durbin procedure results in a lower variance on the variance estimate.

A final advantage is that the underlying Durbin structure allows us to estimate unbiasedly for this survey the increase in variance due to the sample reduction and the variance before the reduction. Thus, we will be able to evaluate the effect of the reduction on our estimates. The variances are currently being estimated.

FOOTNOTES

[1]The Current Population Survey is a monthly survey conducted by the Bureau of the Census for the Bureau of Labor Statistics. Its prime purpose is to produce monthly labor force data, but in March of each year an extensive set of supplementary questions on income and household composition are asked which makes possible estimates of children in poverty families. For more details, see Thompson [9].

[2]The basis for the formulae and methodology used is given in chapter 6 of Hansen, Hurwitz, and Madow [6]. Notation here is not generally consistent with the book.

[3]Strictly speaking, this formula was not used and the resultant data in table 2 was not produced in our earlier work. This is equivalent

to the earlier work, though, and is presented in this form for ease of comparison.

[4]Recall that the final sample was selected from the 20-percent census data.

[5]Page 267, Hansen, et al. [6]

[6]Cochran [3] has shown that if the population is autocorrelated, that is, $\rho_i \geq \rho_{i+1} \geq 0$, and the correlogram is concave upwards, that is $\rho_{i-1}+\rho_{i+1}-2\rho_i \geq 0$, then systematic sampling is superior to stratified sampling taking one unit per stratum, where ρ_i is the correlation between two units which are i units apart in a listing of the population. Because of the sort, described above, that was imposed on the census frame prior to the sample selection, the census population is autocorrelated for estimates of poverty, but no result has been derived that shows the conditions under which a systematic sample is superior to a stratified, two units per stratum without replacement.

REFERENCES

[1] Boisen, Morton. "Study Plan for Sample Design on Survey of Income and Education, Public Law 93-380." Internal Census Bureau memorandum to Daniel B. Levine, May 12, 1975.

[2] Boisen, Morton. "Sample Sizes and Sample Design Decisions for the Survey of Income and Education." Census Bureau memorandum to Wray Smith, Department of Health, Education and Welfare, July 18, 1975.

[3] Cochran, W. G. "Relative Accuracy of Systematic and Stratified Random Samples of a Certain Class of Populations." Annals of Mathematical Statistics, 17, (1946) pp. 164-177.

[4] Dippo, Cathryn. "Expansion of CPS to Provide Reliable State Estimates of Unemployment." Proceedings of the Social Statistics Section, American Statistical Association, 1975, pp. 387-391.

[5] Durbin, J. "Design of Multi-Stage Surveys for the Estimation of Sampling Errors." Applied Statistics, 16, (1967) pp. 52-46.

[6] Hansen, Morris H., Hurwitz, William N., and Madow, William G. Sample Survey Methods and Theory, Vol. 1. New York: John Wiley & Sons, 1953.

[7] Sampford, M. R. "On Sampling Without Replacement with Unequal Probabilities of Selection." Biometrika, 54, (1967) pp. 499-513.

[8] Smith, Wray. "Reduction of the Supplemental Sample for the SIE." Department of Health, Education, and Welfare memorandum to Earle J. Gerson, Census Bureau, Dec. 7, 1975.

[9] Thompson, Marvin M. and Shapiro, Gary. "The Current Population Survey: An Overview." Annals of Economic and Social Measurement, Vol. 2, No. 2, (April 1973) pp. 105-129.

[10] U. S. Congress, Elementary and Secondary Education Act of 1965, Pub. L. 89-10, 89th Cong., 1st sess., 1965, H.R. 2362.

[11] U. S. Congress, Education Amendments of 1974, Pub. L. 93-380, 93rd Cong., 2nd sess., 1974, H.R. 69.

ACKNOWLEDGMENTS

The sample design for this survey resulted from the contribution of many people besides the authors. Of particular importance are Wray Smith, David Bateman, Paul Bettin, Harold Nisselson, and William P. Smith, III. Mr. Bettin also contributed to the writing of Section III of this paper. Helpful comments were made by Wray Smith, Paul Bettin, George H. Gray, Charles D. Jones, Henry Woltman, and an anonymous Census Bureau referee. The typing was done by Edith Oechsler, assisted by Arlene Sagin.

TABLE 3. COSTS AND SAMPLE SIZES FOR ALTERNATIVE NONCOMPACT CLUSTER SIZES

Noncompact Cluster Size	No. of Sample Units Required for 10% CV	No. of Interviews Required Above Minimum (For Cluster Size of 1)	Cost of Training and Recruiting Additional Interviewers	No. of Sample ED's	Variable Cost of Sample Selection	Direct Field Costs	Sum of (6) and (7)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	9,673	--	--	9,673	$34,000	$41,400	$75,400
2	10,343	6	$ 2,100	5,171	18,000	41,305	61,405
3	11,110	14	4,900	3,703	14,000	40,815	59,715
4	11,876	22	7,700	2,969	13,200	41,738	62,638
5	12,644	29	10,150	2,529	13,000	43,092	66,242
6	13,410	37	12,950	2,236	12,000	45,186	70,136

TABLE 4

Between PSU Variance for Selected States for Estimates Based on Stratification by 1970 Poverty Children

	State and Characteristic	2 PSU's per stratum (Durbin)	1 PSU per stratum	Collapse Strata		Number of sample NSR PSU's
				Unadjusted	Adjusted[1]	
		(000)	(000)	(000)	(000)	
a.	Estimated 1970 Poverty Children	(1)	(2)	(3)	(4)	
	Alabama	39,984	18,843	65,671	66,286	12
	California	22,839	10,181	101,697	62,260	8
	Florida	36,182	14,448	64,168	62,866	8
	Michigan	2,380	853	4,167	4,829	12
	South Dakota	2,880	767	4,264	4,245	14
	Washington	918	188	3,940	2,648	8
b.	Estimated 1970 Poverty Families					
	Alabama	11,198	12,639	22,417	27,653	
	California	8,405	7,268	25,835	16,105	
	Florida	23,862	16,075	38,783	37,414	
	Michigan	2,858	2,908	3,174	3,603	
	South Dakota	544	504	818	968	
	Washington	639	448	1,810	1,183	
c.	Estimated 1960 Poverty Families					
	Alabama	44,523	52,535	78,961	96,947	
	California	20,126	22,022	49,029	34,606	
	Florida	61,388	34,726	121,144	107,063	
	Michigan	15,408	15,615	17,844	19,735	
	South Dakota	2,987	2,804	4,576	5,381	
	Washington	1,865	1,922	4,032	3,020	

[1]The adjusted collapsed strata estimate attempts to reduce the bias by eliminating the bias due to the different strata size. For this column, the following estimate was used:

$$(a_1\hat{X}_1 - a_2\hat{X}_2)^2$$

where $a_1 = \frac{2N_2}{N_1+N_2}$ and $a_2 = \frac{2N_1}{N_1+N_2}$,

N_1 and N_2 are the 1970 populations in the two paired strata.

TABLE 5

Bias and Relative Mean Square Error for Selected States in the Adjusted Durbin Estimate of Variance for 1970 Children in Poverty

STATE	Variance Due to the Selection of Primary Sampling Units	Bias in Reducing C_K to 1	Total Variance	Relative Bias Due to Reducing C_K to 1	Relative Variance of the Unbiased Variance Estimate	Relative MSE of the Biased Variance Estimate
	(1) (000)	(2) (000)	(3) (000)	(4) (000)	(5) (000)	(6) (000)
Alabama	39,984	-247	393,801	-0.0006	0.16254	0.15457
Georgia	55,623	-176	638,425	-0.0003	0.19655	0.19189
North Dakota	272	- 90	3,999	-0.0225	0.16848	0.06416
Ohio	4,849	- 5	540,310	-0.0000	0.01798	0.01731
South Dakota	2,880	- 21	12,894	-0.0017	0.23600	0.22448
Utah	945	- 34	6,402	-0.0053	0.07791	0.06808

TABLE 6

Increase in Variance of an Estimate of 1970 Children in Poverty Due to Deleting ED's from the Durbin Housing Unit Sample

STATE	Average Number of Sample HU's Per ED	Approximate Percent Reduction Required	Increase in Variance Over Housing Unit Reduction[1]	Expected Coefficient of Variation from an ED Reduction (%)
	(1)	(2)	(3)	(4)
Maine	2.8	5	1.009	9.4
New Hampshire	NA	2	1.023	9.9
Vermont	5.2	4	1.056	10.1
Rhode Island	NA	5	1.024	9.9
Connecticut	3.1	8	1.006	9.8
Nebraska	3.0	6	1.030	9.8
Kansas	2.9	6	1.025	9.8
Delaware	6.4	20	1.206	10.8
District of Columbia	2.9	10	1.023	9.7
Montana	4.3	7	1.057	9.9
Idaho	NA	5	1.085	10.2
Wyoming	7.7	14	1.770	13.0
New Mexico	3.7	18	1.041	7.0
Nevada	5.7	8	1.136	10.4
Hawaii	5.9	24	1.144	10.5

[1]These were derived from a regression model.