DETERMINING THE OPTIMUM NUMBER OF PRIMARY SAMPLING UNITS TO BE SELECTED FOR THE HEALTH INTERVIEW SURVEY

William H. Tadros, Thomas F. Moore, and R.P. Chakrabarty, U.S. Bureau of the Census

I. INTRODUCTION

The National Health Interview Survey (HIS) is conducted weekly to provide current data on illness and disability and related information of the U. S. civilian noninstitutional population 14 years of age and over. HIS started in July 1957 and has been in operation continuously to the present time.

Research activities for the HIS redesign coincide with the traditional plans formulated by the U. S. Bureau of the Census to redesign its major demographic surveys soon after each Decennial Census of Population and Housing [1]. In the past, HIS and other surveys have shared the design of the Current Population Survey (CPS). For the post 1980 Census redesign, much more consideration is being given to the specific requirements of the individual surveys. Projects for the HIS redesign cover a broad range of topics. One of these is a study to determine the optimum number of sample primary sampling units (PSUs).

The current HIS sample is located in 376 PSUs and only 120 interviewers are assigned to work on the survey. Evidence suggests that the current size of an interviewer's weekly assignment is optimal; increasing it would require either overtime or having interviewers working too many weeks in a year, and decreasing it would not keep the interviewers adequately employed. Because the number of interviewers is less than the number of PSUs, travel to and from nonresident PSUs is extensive and consumes about 35 percent of the direct field costs that cover travel and interviewing. Travel to and from nonresident PSUs could be reduced by decreasing the number of these PSUs. If the reduction is achieved by hiring more interviewers, then a typical interviewer would have too small a yearly workload. A second way to reduce the number of nonresident PSUs is to reduce the total number of PSUs; this is expected to be more cost effective and efficient.

The methodology and results of the study to determine the desired number of PSUs for HIS redesign when cost, precision or sample size is fixed is the topic of this paper.

Initially, simple variance and cost functions were established and results supported the hypothesis that redesigning HIS with fewer PSUs would be more efficient. The initial approach was eventually improved in several ways.

A brief description of the current HIS design is provided in Section II of this paper. The third section gives the strategies used to determine the optimum number of sample PSUs. Results based on the initial model are given in Table 1 and results based on the improved model are presented in Table 2.

II. DESCRIPTION OF THE CURRENT DESIGN

The objective of HIS is to obtain current data pertaining to the health characteristics of the U. S. civilian noninstitutional population 14 years of age and over.

The annual sample of about 52,000 designated housing units for HIS is divided into weekly assignments averaging about 16 to 20 housing units each and is distributed evenly over the 52 weeks of the year. Recent data show that a typical interviewer may have about 26 assignments in a year, or an average of one assignment each 2 weeks. The interviewers are part-time employees and generally work only on HIS.

The selection of sample units takes place in a multistage process. The approximately 3,000 counties, county equivalents and independent cities in the United States are combined into about 1,900 PSUs. A PSU is defined to consist of a Standard Metropolitan Statistical Area (SMSA), a single county, or small group of contiguous counties. All the 1,900 PSUs are stratified into 376 strata with one sample PSU selected from each stratum. Each PSU with population greater than about 250,000 persons in 1970 composes a stratum by itself and enters the sample with certainty. These PSUs are designated as self-representing (SR). Other strata, designated as nonself-representing (NSR), are formed by combining PSUs which are similar in selected characteristics. The current 376 strata design consists of 156 SR PSUs and 220 NSR PSUs. These are the same sample PSUs designated for the 376 PSUs design of the Current Population Survey (CPS).[1]

The remaining selection stages lead to the inclusion of segments in the sample. Segments are clusters of an expected four housing units selected within the sample PSUs. Each household is interviewed only once. Additional aspects of the design are given in [2].

III. STRATEGIES

Initially, we examined the effect of using fewer sample PSUs for redesigning HIS. Variance and cost functions were established with costs based on travel costs to and from nonresident PSUs and on travel and interviewing costs within PSUs. Different assumptions concerning the design parameters were also considered (e.g., different estimates for the proportion of total variance due to variance between PSUs, different segment sizes, ...). Several improvements and refinements were used in the second and final examination. The cost function was modified in that the costs were based on costs per working assignment. The actual travel costs for nonresident PSUs depend on the total annual workload within these PSUs. The initial cost function tended to underestimate costs if the workload in the nonresident PSUs did not decrease at the same rate as the number of nonresident PSUs. The variance function was refined so that it reflected the changes in the number of sample NSR PSUs and populations of the NSR areas. In the earlier model the between PSU variance was treated as being dependent on the total number of PSUs.

The initial examination described under the PSU model is presented in Part A of this section, while the improved method described under the working assignment model is given in Part B.

A. PSU MODEL

The first step in this model involved the establishment of appropriate variance and cost functions.

Variance Function

Generally, the relvariance [2] of an estimated mean or percentage, V_r^2, can be expressed as

$$V_r^2 \doteq V^2 \frac{1}{m\bar{n}\bar{\bar{q}}} [\delta_1 \bar{\bar{n}}\bar{q} + 1 + \delta_2 (\bar{\bar{q}}-1)] . \quad (1)$$

where

V^2 is the population relvariance between housing units within strata;

δ_1 is the average within strata measure of homogeneity between housing units within PSUs;

δ_2 is the average measure of homogeneity between housing units within segments (computed within PSUs);

m is the number of sample PSUs;

$\bar{n}$ is the average number of sample segments per PSU; and

$\bar{\bar{q}}$ is the average number of sample housing units per segment.

Data produced in a study by the Census Bureau [4] show that for many important health characteristics δ_2 ranges from 0.01 to 0.17. We found that the optimal number of sample PSUs is not very sensitive to the value of δ_2. In this paper, the value of δ_2 is fixed at the level of 0.05. The best available variance estimates indicate that the between PSU variance is about 10-20 percent of the total variance in the present design. Three levels of between PSU variance are used in this paper: 10, 15, and 20 percents. Values of δ_1 consistent with these percentages are assumed.

Cost Function

In arriving at the optimum design, consideration was given to how the costs arise. As mentioned before, the HIS sample is located in 376 PSUs, and only 120 interviewers work on HIS. For cost purposes, HIS PSUs are grouped as resident PSUs and nonresident PSUs. A resident PSU is a PSU where at least one interviewer lives. Nonresident PSUs may require travel by privately owned or public transportation (mostly flying) or overnight living expenses (per diem). Thus, the following cost function was established:

$$C = C_{NR}\, m_{NR} + C_{BAL}\, \bar{\bar{m}}\bar{\bar{n}}\bar{\bar{q}} , \quad (2)$$

where

C is the annual amount paid directly to the interviewers for travel, interviewing and related expenses;

C_{NR} is the annual average travel cost per nonresident PSU;

m_{NR} is the number of nonresident PSUs; and

$C_{BAL}\, \bar{\bar{m}}\bar{\bar{n}}\bar{\bar{q}}$ is the balance of costs dependent on the sample size and covers travel within PSUs and interview costs.

The average travel cost to and from a nonresident PSU (C_{NR}) was computed as \$1,191. Also, the average cost for interviewing and travel within PSU (C_{BAL}) was computed as \$13.06 per housing unit for segments of size 4. This average was adjusted when segment size changes from the current size of 4 to 8 or 10. [3]

Later we considered models where the costs for different travel expenses (e.g., air fare, per diem) changed at different rates when the number of PSUs changed, but this had little effect on the optimum numbers of sample PSUs.

Additional aspects of cost constants are provided in [6] and [7].

Results

After establishing the appropriate variance and cost functions as given by (1) and (2), various values of m were assumed and values of V_r^2/V^2 and C computed. In order to give the corresponding value of m_{NR} for each of the selected values of of m, we decided that the number of resident PSUs would remain at the current level of 73. Therefore, $m_{NR} = m - 73$. This assumption is not entirely valid when the number of PSUs decreases significantly. However, several numbers of resident PSUs were examined and we found that when other factors remain unchanged, the optimum number of PSUs remains almost unchanged.

When either cost or precision is fixed, the optimal number of PSUs is the one that minimizes variance or cost, respectively. When the sample size is fixed, the optimal number of PSUs is the one that minimizes the product of cost and variance.

The optimum number of sample PSUs under different conditions are given in Table 1.

B. WORKING ASSIGNMENT MODEL

Variance Function

For our purposes, the between PSU variance is more correctly expressed as a function of m_{NSR}, the number of sample NSR PSUs, rather than the total number of PSUs. To allow for this and to permit separate adjustments to the within and between PSU components of variance for alternative designs, formula (1) was rewritten as

$$V_r^2/V^2 \doteq \frac{\delta_1}{m_{NSR}} + \frac{1+\delta_2 (\bar{\bar{q}}-1)}{\bar{\bar{m}}\bar{\bar{n}}\bar{\bar{q}}} . \quad (3)$$

We divided equation (3) by the quantity V_r^2/V^2 to get

$$1 = V(B) + V(W). \quad (4)$$

The terms V(B) and V(W) are the proportions of variance between and within PSUs, respectively, for the present design.

In the designs under consideration, the between PSU variance would be affected by including a much larger NSR population in the new designs, and by an improved procedure which has been developed for stratifying PSUs. Experimental results from separate work showed that the combined effect of these influences was an increase of about

3 percent in the variance. Thus, the between PSU variance of a new design relative to the current design was calculated using

$$V(B) \times \frac{220}{m_{NSR}} \times 1.03 , \qquad (5)$$

where 220 is the current number of NSR PSUs and m_{NSR} is the number of NSR PSUs in a new design.

The within PSU component was adjusted for alternative designs by allowing for changes in the sample size and segment size. The within PSU variance of a new design relative to the current design was given as:

$$V(W) \times \frac{51,755}{\bar{m}\bar{n}\bar{q}} \times \frac{1 + \delta_2(\bar{q} - 1)}{1 + \delta_2(4 - 1)} , \qquad (6)$$

where 51,755 is the current sample size, 4 is the present segment size, $\bar{q}$ is the new segment size, and $\delta_2 = .05$, a typical value for the intraclass correlation coefficient. Thus, the relative variance, f_v, of a new design was given as:

$$f_v = \frac{\text{variance of a new design}}{\text{variance of the current design}} = \frac{(5) + (6)}{(4)}$$

$$= \frac{V(B) \times 220 \times 1.03}{m_{NSR}} + \frac{V(W) \times 51,755 \times [1+\delta_2(\bar{q}-1)]}{\bar{m}\bar{n}\bar{q} \times [1+\delta_2(4-1)]} . \qquad (7)$$

Cost Function

The cost function (2) used in the PSU model assummed that the travel cost to and from nonresident PSUs depended on the number of nonresident PSUs. The actual travel costs for nonresident PSUs also depend on the total annual workload within these PSUs. Formula (2) would tend to underestimate costs of a new design if the workload in these PSUs did not decrease at the same rate as the number of PSUs. In order to improve cost estimates, a new cost function was established based on costs per working assignment.[4] The improved cost function was

$$C = C_R W_R + C_{NR} W_{NR} , \qquad (8)$$

where

C_R and C_{NR} are the per working assignment costs dependent on the type of working assignment; and

W_R and W_{NR} are the numbers of resident and nonresident working assignments, respectively. A nonresident working assignment is one that is in a nonresident PSU.

Values [5] for C_R were \$209 for segments of size 4 and \$167 for segments of size 8 or 10. The additional cost for travel to and from nonresident working assignment was \$230 regardless of the segment size and, therefore, C_{NR} was \$439 for segments of size 4 and \$397 for segments of size 8 or 10. Additional aspects of costs are given [9].

The relative cost, f_c, for a new design was computed as:

$$f_c = \frac{\text{cost of a new design}}{\text{cost of the current design}} . \qquad (9)$$

Results

The optimal number of sample PSUs for a given cost, variance or sample size was determined in the following manner. If we fixed the cost, then for each combination of resident and nonresident working assignments that would give us that cost, we would find a combination of SR and NSR PSUs that was consistent with the mix of assignments and compute the variance. If we fixed the variance, we would find combinations of SR and NSR PSUs and sample size to give that variance and, after allocating sample and interviewers to the PSUs, compute the cost. For a fixed sample size, we would select combinations of SR and NSR PSUs, allocate working assignments and interviewers to PSUs, and find the number of nonresident working assignments. See [9] for additional details. Then for each combination we computed f_v and f_c.

We attempted to find combinations of SR and NSR PSUs so that the smallest SR PSU would have about the same population as the average NSR stratum. Populations of SR PSUs were determined by using a list of Standard Metropolitan Statistical Areas (SMSAs), and counties if necessary, ranked by the 1980 census population. Working assignments and interviewers were allocated to SR and NSR PSUs in proportion to stratum populations.

As in the PSU model, when either cost or precision is fixed, the optimal number of PSUs is the number that minimizes f_v or f_c, respectively. When the sample size is fixed, the optimal number of PSUs is the one that minimizes $f_v \times f_c$.

IV. DISCUSSION

Descriptions of the optimal designs under the working assignment model are provided in Table 2. Results support the hypothesis that redesigning HIS with fewer PSUs would be more efficient. For example, if cost, variance or sample size is fixed at the current level, then the optimum number of sample PSUs would be in a 104-174 range when segment size is increased from the current level of 4 to 8. With the optimum number of PSUs and the indicated sample sizes, a reduction in variance (cost) is expected when cost (variance) is fixed at the current level. When sample size is fixed at the current level of 52,000 housing units, a reduction in costs is accompanied by an increase in variance.

We found that designs where the annual interviewer workload in NSR PSUs was about large enough to support a single interviewer tend to be optimal. This meant that there would be no nonresident working assignments. With less travel, a larger sample size is possible when cost is fixed. A larger sample size produces a smaller within PSU variance which more than compensates for the increase in the between PSU variance that would result from the decrease in the number of sample NSR PSUs.

Also, we found that when costs are fixed at the current level of \$1,037,000 [6] the sample sizes of the optimal designs, when segment size is 8 or 10, are almost twice the current sample size (99,400 vs 52,000). The current costs support 1,664 resi-

dent working assignments and 1,571 nonresident working assignments at an average of $209 and $439 per resident and nonresident working assignment, respectively. As mentioned before, an optimal design would include no nonresident working assignments and, therefore, the sample size of an optimal design when segment size is 4 will be about 79,400. Also, the cost of a resident working assignment decreases from $209 to $167 when segment size increases from 4 to 8 or 10, and, as a result, the sample size of an optimal design would increase to about 99,400.

Table 2 shows that optimal designs would require more interviewers than the current number of 120 when cost or variance is fixed. Although the increase in the number of interviewers would eliminate the costs of travel to and from nonresident working assignments, there would still remain the costs of recruiting and training the newly hired interviewers. Such costs are not included in the work presented in this paper.

Several factors could change the optimum. For example, the best available variance estimates indicate that the between PSU variance is about 10-20 percent of total variance in the current design. If new estimates pointed to a higher percentage, we might want to add PSUs. If we dropped the requirement that interviewers be kept occupied on HIS for about 26 weeks a year, we could probably add PSUs and still have a minimum travel design. Finally, other factors may influence the decision on the number of sample PSUs. One such factor may be a need for analytical capability at subnational levels such as medically underserved areas of the country.

V. ACKNOWLEDGEMENTS

Special thanks are due to Edith Oechsler for typing several drafts and to Linda Sinanian for typing the final copy of this paper.

* * * * * * * *

FOOTNOTES

[1] Some small PSUs are not large enough to supply the sample needed for all surveys using the 376 PSUs design and must be replaced over time. The replacement schedule differs from one survey to another and, therefore, the 376 PSUs in HIS may be slightly different from the 376 PSUs in the CPS sample [2].

[2] See [3], V.I., page 370.

[3] This adjustment was based on a unit cost study done for HIS in March 1971. See [5].

[4] Currently the average size of a working assignment is 16-20 housing units.

[5] Simultaneously with this work, the cost study in [6] was being updated (see [8]). We were able to use these updated cost estimates in this model. The optimum numbers of PSUs were not recalculated for the PSU model.

[6] Actual field costs during fiscal year 1980.

REFERENCES

[1] Kniceley, R. Maurice, and Baer, Leonard R., "General Overview of Research in Redesign of the Census Bureau's Demographic Surveys," unpublished Census Bureau report, March 1982. (A shorter version of this paper, with the same title, was published in the Proceedings of the Section on Survey Research Methods, 1981 American Statistical Association Meetings, pp. 214-219).

[2] U.S. Department of Commerce, Bureau of the Census, The Current Population Survey: Design and Methodology, Technical Paper 40, Washington, D.C. Government Printing Office, 1978. Written by Robert H. Hanson.

[3] Hanson, M.H., Hurwitz, W.N., and Madow, W.G. 1953, Sample Survey Methods and Theory, John Wiley and Sons, New York, Vol. I.

[4] U.S. Department of Health, Education, and Welfare, National Center for Health Statistics, Reliability of Estimates with Alternative Cluster Sizes in the Health Interview Survey, Washington, D.C., Government Printing Office, 1973.

[5] U.S. Department of Commerce, Bureau of the Census, "Optimal Segment Size for Health Interview Survey," memorandum from Joseph Waksberg to Monroe Sirken (NCHS), March 2, 1971.

[6] U.S. Department of Commerce, Bureau of the Census, "HIS Redesign-Description of Constants of Variance and Cost Functions Needed to Determine the Desired Number of Sample PSUs," memorandum from Rameswar P. Chakrabarty to Gary M. Shapiro, written by William Tadros, September 2, 1981.

[7] U.S. Department of Commerce, Bureau of the Census, "HIS Redesign-Determining the Optimum Number of PSUs," memorandum from Rameswar P. Chakrabarty to Gary M. Shapiro, written by William Tadros, September 14, 1981.

[8] U.S. Department of Commerce, Bureau of the Census, "Report on Optimum Segment Size Study for Health Interview Survey (HIS)," memorandum from Elizabeth T. Huang to Gary M. Shapiro, May 4, 1982.

[9] U.S. Department of Commerce, Bureau of the Census, "HIS Redesign-Determining the Optimum Number of Sample PSUs, Working Assignment Model," memorandum from Rameswar P. Chakrabarty to Gary M. Shapiro, written by William Tadros, September 3, 1982.

Table 1. Optimal Numbers of Primary Sampling Units

for HIS Redesign [δ_2 = .05]
(based on the PSU model)

Percent Between Variance P	Segment Size = q	Optimum Number of PSUs that provide the Minimum Variance for Fixed Cost	Optimum Number of PSUs that provide the Minimum Cost for Fixed Variance	Optimum Number of PSUs that provide the best combination of cost and variance for a sample size of[1] 50,000 Housing Units	45,000 Housing Units	37,500 Housing Units
10	4	200 (66,100)	174 (57,400)	141	126	103
	8	197 (67,600)	168 (58,000)	135	120	98
	10	196 (68,100)	167 (58,100)	133	119	97
15	4	239 (62,500)	219 (57,300)	178	158	130
	8	234 (64,000)	212 (57,900)	169	151	123
	10	234 (64,300)	211 (58,000)	168	149	122
20	4	271 (59,600)	257 (56,500)	211	189	154
	8	268 (60,900)	252 (57,200)	204	182	149
	10	265 (61,200)	248 (57,400)	200	178	145

[1]Reduces cost but increases variance compared to the current design.

NOTES: A. Numbers in parentheses provide the sample size.

B. Any number within about 20 PSUs of the optimum would be about as good as the optimum.

TABLE 2. Optimal Numbers of PSUs for HIS Redesign under Listed Conditions [δ_2=0.05]
(based on the working assignment model)

Conditions	P = .10			P = .15			P = .20		
	Segment Size ($\bar{q}$)			Segment Size ($\bar{q}$)			Segment Size ($\bar{q}$)		
	4	8	10	4	8	10	4	8	10
1. Minimum Variance for Fixed Cost									
Number of PSUs	146	174	174	146	174	174	146	174	174
Number of SR PSUs	32	38	38	32	38	38	32	38	38
Percent of Population in SR PSUs	.38	.41	.41	.38	.41	.41	.38	.41	.41
Sample Size	79,400	99,400	99,400	79,400	99,400	99,400	79,400	99,400	99,400
Design's Variance / Current Variance	.79	.72	.76	.85	.77	.81	.92	.82	.86
2. Minimum Cost for Fixed Variance									
Number of PSUs	118	132	138	128	140	145	135	145	153
Number of SR PSUs	22	27	29	26	30	32	28	32	34
Percent of Population in SR PSUs	.32	.35	.36	.35	.37	.38	.36	.38	.39
Sample Size	61,000	69,700	74,200	66,700	74,800	79,400	71,400	81,200	84,200
Design's Cost / Current Cost	.77	.70	.75	.84	.75	.80	.90	.82	.85
3. Best Combination of Cost and Variance[1]									
a. For a sample size of 52,000									
Number of PSUs	104	104	104	104	104	104	104	104	104
Number of SR PSUs	19	19	19	19	19	19	19	19	19
Percent of Population in SR PSUs	.30	.30	.30	.30	.30	.30	.30	.30	.30
Design's Cost / Current Cost	.66	.52	.52	.66	.52	.52	.66	.52	.52
Design's Variance / Current Variance	1.16	1.32	1.40	1.25	1.39	1.47	1.33	1.47	1.54

[1]It may be noted that for a fixed sample size, the optimum number of PSUs reduces the cost but increases the variance.

P = Proportion of between PSU variance to total variance.

NOTE: Number of interviewers could be derived from (sample size ÷ 431)
where 431 is the current yearly workload in housing units of a typical interviewer.