THE COST OF SAMPLE SURVEY DESIGNS

Charles P. Humphreys, World Bank[1]

1. INTRODUCTION

Methods to analyze and reduce the costs surveys must receive greater attention if research budgets are to be utilized efficiently. While it is impossible here to consider all factors that contribute to the unit cost of information, this paper analyzes the effects of the design and size of sample surveys. In particular, the analysis focuses on a two-stage, cluster sample design. The framework is general, but the empirical assessment relies heavily on the design and cost structure of the Ada Baseline Survey (ABS), conducted in rural Ethiopia during 1973. In addition, some conclusions emerge that provide guidelines for planning similar surveys.

The paper contains five major sections. The following section develops an algebraic cost function that can be used to assess the cost-effectiveness of cluster sampling designs. It is followed by two sections that describe the ABS and estimate the coefficients of the cost function using cost data from that study. The last two sections assess the ABS design and offer guidelines for reducing survey costs without sacrificing precision.

2. COST FUNCTION

A fully generalized cost function draws upon differential calculus and decision theory to minimize variances subject to a given cost (see Cochran, 1963, pp. 82-84), but a simpler algebraic relationship permits an examination of the effects of survey design on survey costs. The costs of carrying out a two-stage cluster sample survey can be divided into five major categories: fixed support costs, travel among first-stage clusters, sampling costs associated with clusters, travel within the clusters, and sampling costs associated with the units of observation. Therefore, an analytical cost function may be defined as:

$$c = c_0 + c_1\sqrt{n} + c_2 n + c_3 n\sqrt{\bar{m}} + c_4 n\bar{m} \qquad (2.1)$$

where n is the number of first-stage clusters in the sample; $\bar{m}$ is the average number of second-stage sampling elements per cluster; c_0 represents total fixed support costs (overhead); c_1 represents the cost coefficient for average travel among first-stage clusters (intercluster travel costs); c_2 represents the average cost coefficient for using a cluster as a sampling unit; c_3 represents the cost coefficient for average travel within each selected locality to contact second-stage sampling elements (intracluster travel costs); and c_4 represents the average cost coefficient for surveying a second-stage sampling element.

The fixed support costs (c_0) mainly comprise the expenses of preliminary studies to design the survey and plan the field work, the contribution by the central research staff to the preparation and pretesting of questionnaires, and secretarial and clerical support by the central office during the survey. These costs can be considered relatively independent of moderate variations in sample size.

Travel costs include both the cost of the transport itself (such as fuel and the maintenance and depreciation of vehicles) and the time that field personnel must spend traveling. The costs of personnel time spent in travel are by far the most important. Although vehicle expenses and travel time are a function of travel distances, other factors--including population density and the physical accessibility of the survey region--will also have an impact. Travel distance is a function of the physical size of the survey region, the sample size, the extent of recalls and replacements, and the clustering of the sample design.[2]

Costs associated with the non-travel activities of the survey include drawing samples, preparing survey materials, locating, identifying and interviewing respondents, and coding data. Material costs are a minor share of such expenses, and include the costs of the survey questionnaire, materials to store the data (computer cards and tapes), and incidentals like maps, clipboards, and pencils. The major costs are for personnel. Personnel costs comprise compensation and indemnities, augmented to account for their training and field supervision (see Hansen, Hurwitz, and Madow, 1953, pp. 285-93). Four major factors affect these personnel costs: sample size (the number of interviews), initially locating and identifying respondents in the sample, the frequency of recalls and replacements, and the duration of each interview. In addition, the method of coding questionnaires will affect personnel costs necessary to translate data into a form ready for analysis.

The assessment of survey designs requires some additional reorganization of terms in (2.1). First, the average survey cost per household is based on the intracluster travel cost coefficient (c_3) and the average household sampling cost coefficient (c_4). This new cost term is:

$$c_5 = \frac{c_3 n\sqrt{\bar{m}} + c_4 n\bar{m}}{n\bar{m}} \qquad (2.2)$$

The effect of the combination means that intracluster travel costs may be understated, but in small clusters the distortion is not great. Second, the total variable survey cost is defined as:

$$c' = c - c_0 \qquad (2.3)$$

3. ADA BASELINE SURVEY

This section describes the Ada Baseline Survey (ABS), which was part of the initial study of a rural area that was to become the focus of an agricultural development project. The purpose of the survey was to provide economic, agricultural, and sociological information that could both assist in the implementation of the project and provide the basis for its subsequent evaluation.

The survey area is a rectangle that covers about 1,750 square kilometers and at the time of the survey contained approximately 21,500 rural households. Of the estimated 108,000 persons in the rural areas, roughly 80 percent were farmers. The region was divided into 229 localities, based largely on traditional judicial circuits, which ranged in size from 19 to 432 households, with an average of about 95. The large urban regional capital was excluded from the sampling frame.

The design of the survey was based on a two-stage cluster sampling procedure (Humphreys, 1974). The first-stage sample consisted of 87 administrative localities (n), giving a first-stage sampling fraction of 38 percent. This sample was systematically selected with probabilities proportional relative to cluster size, where size is defined by the number of households resident in the cluster. By drawing the first-stage sample with probabilities proportional to size, the second-stage sample could consist of approximately equal numbers of farm households randomly selected (with replacement) from each locality in the first-stage sample (Hartley and Rao, 1962, and Cochran, 1963). The final second-stage sample contained 632 observations. This sample size reflects adjustments for revised population estimates in some localities and for households that could be neither located nor replaced. This sample size gives an average of 7.26 households per locality ($\bar{m}$) and an overall sampling fraction of 2.9 percent.

This particular design was selected because it offered several advantages. Systematic sampling (without replacement) at the first stage allowed the selection of a large number of localities, which assured extensive coverage of the geographically heterogeneous region. Secondly, because the sample is self-weighting, second-stage samples were of similar size in each locality, which facilitated the logistical organization of the field work. Thirdly, the two-stage design permitted interviewing to begin before the entire sample was drawn, thereby easing the constraint on available time.

The field work was carried out by two survey teams, each composed on average of 7 enumerators, a supervisor, a driver, and one vehicle. The survey required 80 team-days of field work, and interviews averaged about 1.5 hours each, including introductions and formalities. Coding demanded an additional month by a group of 5 coders. On average, the round-trip distance between the centrally located field office and the first-stage sampling clusters was about 60 km. This travel consumed roughly 30 percent of the working day of the survey teams. In addition, about one-sixth of the time in the field was spent in becoming familiar with the sample localities, including their location, access, and residents.

Locating and identifying sample households proved more difficult than anticipated despite the clustering, offseting some of the expected advantages of the design. The considerable time required to travel to and become familiar with localities, the high rate of replacement of households originally selected (about 25 percent), and the high absentee rate (estimated at about 40 percent of the first visit and about 25 percent on subsequent visits) meant that each enumerator averaged only 1.25 interviews per day. As a consequence, each locality had to be visited several times, which increased travel time and expense, reduced enumerator productivity, and prolonged the field work.

4. EMPIRICAL ESTIMATES

Estimates of the cost coefficients are made by dividing the total costs for each term in the cost function by the appropriate expression for the relevant sample size. Values for total costs --based on the ABS--appear in Table 1.

1. Survey Costs

Cost Categories	US$	Percent
Fixed Costs (c_0)	3,589	24
Preliminary Planning	1,813	51
Questionnaire Development	1,335	37
First Stage Costs	5,547	37
Intercluster Travel ($c_1\sqrt{n}$)	2,342	42
Vehicle	504	21
Personnel	979	42
Surcharges	859	37
Cluster Sampling ($c_2 n$)	3,205	58
Materials and Travel	561	17
Personnel	1,791	56
Surcharges	853	27
Second Stage Costs	5,899	39
Intracluster Travel ($c_3 n\sqrt{\bar{m}}$)	2,281	39
Vehicle	679	30
Personnel	857	37
Surcharges	745	33
Element Sampling ($c_4 n\bar{m}$)	3,618	61
Materials	337	9
Personnel	2,276	63
Surcharges	1,005	28
Total Costs	15,035	100

Survey costs for the first-stage cluster consist of intercluster travel costs and cluster sampling costs. Although total travel costs vary mainly with the square root of first-stage sample size--assuming the areas of the clusters are similar, the cost coefficient itself is influenced by a combination of multiple recalls and a large number of enumerators per vehicle, which resulted in a zigzag pattern of cross-regional travel rather than in a systematic progression through neighboring localities. Intercluster travel costs represented about two-fifths of total costs associated with the first-stage clusters. Of these travel costs, vehicle operation accounted for less than one-quarter, costs related to personnel accounting for the remaining three-quarters. However, direct personnel costs for actual field travel amounted to less than half of travel costs, while surcharges--including overheads for enumerator training and central staff support as well as those caused by inefficient field operations--equaled nearly 40 percent of these costs.

Cluster sampling costs cover all the expenses associated with defining, selecting, and utilizing the cluster as a first-stage sampling element.

Such costs include the time required to become familiar with each selected cluster, e.g., physically locating it, establishing rapport with the community, and learning the pattern of roads within it. The magnitude of these costs was more important than travel costs, and costs associated with personnel were a much larger share of the total.

Second-stage costs are analogous to cluster survey costs, consisting of intracluster travel costs and element sampling costs. In principle, the travel costs within clusters vary with the square root of the average second-stage sample size, but these costs are increased by initial and subsequent callbacks, the inability to locate households, and sub-optimal field organization. As with cluster survey costs, travel costs amounts to about two-fifths of total household survey costs. But vehicle operation was more important, accounting for about one-third, while direct personnel costs were less than 40 percent.

Element sampling costs are those associated with the actual interviewing process and vary directly with the final sample size. The single most important cost is enumerator time for interviewing respondents, although it accounts for only one-quarter of the total. All direct interviewing costs amount to nearly 40 percent of second-stage costs. The remaining costs consist of personnel surcharges, data coding, and data storage, in order of importance. Material costs in all categories were relatively unimportant, amounting to only 9 percent of all element sampling costs.

In summary, the field costs (variable expenses) amounted to about three-fourths of the total survey costs. Field costs were almost equally divided between those associated with the first-stage cluster sample and the second-stage household sample. Of the field costs, direct and indirect travel costs constituted almost half, although vehicle operation comprised less than one-fourth of these travel costs. Total personnel costs--including surcharges--amounted to over 80 percent of all field costs, although the share drops to only three-fifths if only direct personnel costs are considered (including the travel time but excluding the surcharges for training and supervision by the central staff). By contrast, material costs made up less than 10 percent of the field costs. Although almost one-half of the direct personnel costs for the field work were for enumerators and coders, only about 15 percent were for actual household interviewing. In fact, of total personnel costs, interviewing took less than 10 percent, and of total field work costs, interviewing consumed only one-twelfth.

On the basis of these costs and actual sample sizes, the values for the cost coefficients can be estimated and are shown in Table 2.

2. Cost Coefficients

Coefficients	Values (US$)
c' (total variable costs)	11,445
c_0 (total fixed costs)	3,589
c_1 (intercluster travel costs)	251
c_2 (cluster sampling costs)	37
c_3 (intracluster travel costs)	10
c_4 (element sampling costs)	6
c_5 (average element costs)	9

5. ASSESSMENT OF SURVEY DESIGN

This analysis assumes two stages and that the definition of the boundaries and sizes of clusters is based on existing administrative units. Therefore, only three features of the design are assessed: the sampling procedure for selecting the first-stage sample of localities, the first-stage sample size (number of clusters, or n), and the average second-stage sample size per cluster (number of elements surveyed in each selected locality, or $\bar{m}$).

There are three basic sampling methods for selecting the first stage sample: sampling with probabilities proportional to relative cluster size--as in the ABS, where cluster size is defined by population, not area; sampling with equal probabilities for all clusters, as in a simple random sample; and sampling with probabilities based on the relative magnitudes of the square roots of the cluster sizes, this method being a compromise between the first two. The choice among these methods depends largely on the ratio of the intercluster travel coefficient (c_1) to the cluster sampling coefficient (c_2) (see Cochran, 1963). If the two coefficients are similar, the third method is preferred. The second method is favored if c_2 is the larger value. And if c_1 is the larger, the first is recommended. For the ABS, c_1 is almost six times larger than c_2 which indicates that some method of sampling with probability proportional to size is the most appropriate first-stage sampling method.

The choice of optimal sample size is more complex because it requires information about both survey costs and cluster homogeneity. Because the first- and second-stage samples are interrelated, the process is iterative. The first step is determining the optimal average second-stage sample size per cluster, which can then be used together with cost information to determine the first-stage sample size. This new first-stage sample size is used to recompute the optimal second-stage sample size.

Given the cost function (2.1), the optimal average second-stage sample size per cluster is:

$$\bar{m}_{opt} = \sqrt{\left(\frac{1-\rho}{\rho}\right)\frac{c_2 + c_1/a}{c_5}} \qquad (5.1)$$

where ρ is the measure of cluster homogeneity, or intraclass correlation coefficient (see Haggard, 1958), and

where $$a = \frac{\sqrt{1 + 4(c'/c_1)b - 1}}{b}$$

and $$b = \frac{c_2 + c_5 \bar{m}_{opt}}{c_1}$$

Then, the optimal first-stage sample size is:

$$n_{opt} = a^2/4 \qquad (5.2)$$

The resulting sample sizes should give the lowest sampling error for the assumed cost structure (see Hansen, Hurwitz, and Madow, 1953).[3]

In general, the coefficient for intercluster travel (c_1) is relatively unimportant in determining the first-stage sample size because this term increases more slowly than the increase in the first-stage sample size. This term also has relatively minor impact on the average second-stage sample size per cluster, especially if the ratio of c_1 to total variable field costs (c') is low, as in the ABS (0.02). The more important coefficients are those for cluster sample costs (c_2) and the average element survey costs (c_5) (Hansen, Hurwitz, and Madow, 1953, p. 299).

If ρ=0.15 is taken as representative value for cluster homogeneity,[4] then the ABS cost structure implies that a more efficient design would be a total sample size of about 550 households for each variable observed, with an average second-stage sample size of approximately 5.5 households per cluster, distributed over about 100 first-stage localities. Compared to the actual survey design, these results suggest an increase in the first-stage sample size of about 15 percent. The indicated optimal overall sample size is somewhat lower than the actual total sample size of 632 rural households, although it is larger than the actual sample of farm households (493). The optimal average second-stage cluster sample would be less than the actual survey size by about 25 percent.

Optimization is difficult, however, when some respondents do not engage in all activities surveyed, thereby reducing effective sample sizes for those variables. In such cases, it may be necessary to increase total sample size to assure that these variables are estimated with the desired precision. Moreover, if cluster homogeneity is not the same for all variables, the allocation of resources between first and second stages will not be optimal for all variables. Of 11 variables analyzed for the ABS, 7 required a larger first stage sample and 8 a larger total sample, in part because some variables were not relevant for all respondents. Only 5 of the variables required a larger second-stage sample to improve efficiency.

It is useful to contrast this assessment with those based on cost considerations alone. The very high intercluster travel cost coefficient (c_1) would suggest a small first-stage sample, and the high cost coefficient for intracluster travel (c_3) would call for a larger average second-stage sample per locality. The low cost coefficient for household sampling (c_4) would also make a larger second-stage sample feasible. However, the low cost coefficient for cluster sampling (c_2) would argue for a large first-stage sample, especially since c_2 is more important than c_1 in determining the first-stage sample size. When variance conditions are considered, the optimal survey design would have a larger first-stage sample and a smaller second-stage than in the ABS, in spite of the high travel costs.

If survey procedures were made more efficient, such savings would also change the survey design. By improving the field organization and increasing the tolerance for nonrespondents, nonoptimal travel patterns could be reduced. A realistic reduction of c_1 might be about 50 percent, to US$125. In contrast, cluster sampling costs for the ABS might be abnormally low because of previous work in the area. If it had been necessary to define and measure clusters and to prepare lists of households in selected localities, c_2 might have doubled to about US$79. Household survey costs could have been reduced by decreasing the intracluster travel costs (c_3) through changes in the organization of the field work (for example, making each enumerator responsible for his own travel). Benefits may also accrue from lowering household sampling costs (c_4) by reducing the data coding and storage costs (for example, using more pre-coding and replacing keypunchers and electronic scanners). Such changes might realistically reduce c_5 by about one-fifth to US$7. Finally, the fixed costs (c_o) were understated in the ABS because of insufficient questionnaire pretesting and enumerator training, but higher fixed costs would probably have little impact on the relative sizes of the coefficients.

Assuming a cluster homogeneity of ρ=0.15, this improved cost structure indicates a first stage sample of 79 clusters and an average second stage sample size per cluster of 7.97, giving a total overall sample of about 624 households for the same budget. These empirical results confirm that cluster traveling costs (c_1) are relatively less important in survey design, but that cluster sampling costs (c_2) are critical in choosing sample size. Furthermore, the average second-stage sample size seems very sensitive to the average cost per element (c_5).

6. GUIDELINES

Guidelines to improve the cost-effectiveness of sample surveys should focus largely on changes in the research design to conform more closely with cost conditions and population variances. The major alternative to the cluster design described above is stratification. Before stratified sampling can commence, it requires extensive knowledge about the entire population as well as complete and informative lists, which tend to increase both fixed costs (c_o) and the non-travel variable cost coefficients (c_2 and c_4). When these stratification costs are excessive--as is likely in rural areas of less developed countries, clustering offers a means of sampling with a smaller budget. But as it also increases sampling variance, the important guidelines are those that can lower cluster sampling costs and allow a larger sample to offset the increase in variance.

If cluster homogeneity is high (as in the ABS, where ρ=0.15), its impact on the sampling

variance can be reduced by using a large first-stage sample. If the ratio of first-stage to second-stage cost coefficients is relatively low (as in the ABS, where $c_1/c_3 = 26$, $c_1/c_5 = 27$, $c_2/c_4 = 6$, and $c_2/c_5 = 4$), savings will result by shifting survey resources from the second to the first stage. Moreover, the effect of higher cluster homogeneity is further reduced by using a smaller average second-stage sample size. The effect of such optimization is to put 52 percent of variable survey costs in the first-stage, compared to 48 percent for actual survey.

Where clusters are defined geographically, variables affected by ecological, spatial, and institutional factors are usually fairly constant within a cluster but may vary greatly among clusters. As a result, the intraclass correlation coefficient is high, which raises the variance of the sample and reduces the efficiency of survey resources. Since these conditions are likely to exist for agricultural, economic, and sociological surveys--especially in isolated rural areas--efficient cluster sampling will require relatively large first-stage samples and relatively small average second-stage samples. However, if first-stage cost coefficients are extremely high relative to second-stage cost coefficients (if, for example, $c_1/c_5 = 400$ or $c_2/c_5 = 8$), a large first-stage sample may be less cost-effective.

A more subtle change in research design would be a higher tolerance for non-respondents. In order to hold the share of non-respondents in the ABS to only 2.3 percent, total variable costs were increased by perhaps as much as 5-12 percent because of recalls. Methods to reduce absenteeism could reduce the number of non-respondents but might increase some of the cost coefficients. Changes in the organization of the field work offer opportunities to reduce costs by reducing the travel time of the field staff (cost coefficients c_1 and c_3). First, daily travel between the field residence and clusters is estimated to have increased total variable costs by as much as 14 percent. Such costs could be reduced by using mobile teams that do not have a permanent field residence. Second, the use of one vehicle to deliver and pick up each member of a team of enumerators meant that considerable enumerator time was lost in waiting, the value of which may have been as much as 7-16 percent of variable costs. Smaller teams or individual transport (like bicycles) for each enumerator could have reduced this cost. Finally, non-optimal travel patterns may have increased variable costs, exclusive of overhead, by 13-28 percent. An improvement in travel patterns (to allow an orderly progression from cluster to cluster), could reduce these costs. Reductions in recalls and individual enumerator travel would also improve travel patterns.

A number of obvious measures have little impact on overall costs. Shorter interviews would save enumerator time, but time actually spent in interviews is only 6 percent of total costs. In fact, because simply reaching respondents is so costly, longer, better interviews may be desirable. Methods to facilitate data handling could reduce coding costs, but these expenses amount to only 4 percent of total costs. Lastly, reductions in the cost of materials would allow only relatively minor savings, since materials amount to only 4 percent of total costs.

In conclusion, the greatest scope for cost savings is in the reduction of travel time. However, efficient sampling designs that offer both low cost and low variance would seem to require a large number of clusters, small samples within clusters, and a highly efficient field organization composed, in part, of small, highly mobile, and self-sufficient interview teams.

APPENDIX - CALCULATION OF FIELD TRAVEL COSTS

Estimates of the cost function coefficients c_1 and c_2 in (2.1) are given by:

$$c_j = x(1 + \sum_{k=1}^{K} i_{ck})(c_p + c_v)\sqrt{A} \qquad (A.1)$$

where c_j is the j^{th} coefficient in (2.1) for travel costs; A is the size of the survey area in square kilometers; c_v is vehicle cost per kilometer; c_p is the value of the time of the survey team personnel spent in traveling; i_{ck} are factors which increase costs by increasing the effective number of sample units; and x is a factor to estimate increases in travel distance and time caused by non-optimal travel patterns. The most important factors increasing costs are initial callbacks because of absenteeism (i_{c1}); additional callbacks for continued absenteeism (i_{c2}); replacement of sample elements determined to be missing (i_{c3}); and returns to pickup enumerators on completion of their interviews (i_{c4}). Finally, before this result can be used in (2.1), personnel costs must be augmented to account for training, central staff support, and travel to and from clusters. Values for these factors in the ABS appear in Table 3. The overhead rate for enumerator training was 4.8 percent of the field costs of enumerators; the surcharge for central staff support amounted to 42 percent of total field personnel costs; and the extra cost of traveling to and from clusters was estimated at 43 percent of direct personnel costs.

3. Travel Cost Factors

Factors		Values	
		c_1	c_3
A	(square kilometers (km)	1750.	7.65
c_v	(US$/km for vehicles)	0.16	0.16
c_p	(US$/km for personnel time spent in travelling)	0.27	0.24
i_{c1}	(initial callbacks for absenteeism)	1.0	0.4
i_{c2}	(subsequent callbacks for absenteeism)	0.0	0.25
i_{c3}	(replacement of missing sample elements)	0.0	0.25
i_{c4}	(returns for enumerator pickups)	1.5	1.0
x	(adjustment for non-optimal travel patterns)	2.5	2.0

FOOTNOTES

[1]Support for this paper has also been provided by the Fletcher School of Law and Diplomacy, Tufts University, the Department of Agricultural Economics, Michigan State University, and the Food Research Institute, Stanford University.

Although the author is a staff member of the World Bank, the views expressed here are his and not necessarily those of the World Bank.

[2]The method for estimating c_1 and c_3 is explained in the Appendix.

[3]Cochran (1963, pp. 279-83) suggests that a more abbreviated estimate of $\bar{m}_{opt}$ can be used when intercluster travel costs are low.

[4]This is the median value of the third quartile for the range of ρ calculated for 36 variables from the ABS (see Hansen, Horwitz, and Madow (1953), p. 407).

REFERENCES

Cochran, William G. (1963), Sampling Techniques (2nd Ed.), New York: John Wiley and Sons, Inc.

Haggard, Ernest A. (1958), Intraclass Correlation and the Analysis of Variance, New York: The Dryden Press, Inc.

Hansen, Morris H., William N. Horwitz, and William G. Madow (1953), Sample Survey Methods and Theory, Vol. 1, New York: John Wiley and Sons, Inc.

Hartley, H.O., and J.N.K. Rao (1962), "Sampling with Unequal Probabilities and Without Replacement," Annals of Mathematics and Statistics, 33, 350-74.

Humphreys, Charles P. (1974), "Ada Baseline Survey: Technical Appendix," Working Paper No. 10, Part II, Addis Ababa: Institute of Development Research, University of Ethiopia.

4. Analysis of ABS Sample Sizes

Variable	ρ	n	n_{opt}	$\bar{m}$	$\bar{m}_{opt}$	$n\bar{m}$	$(n\bar{m})_{opt}$	Decrease in Variance
Household Size	0.05	87	72	7.26	9.95	632	716	3
Organizational Membership	0.16	87	103	7.26	5.28	632	544	2
Cropland per Farm	0.06	83	75	5.80	9.37	481	703	27
Crop Value per Farm	0.14	82	99	5.73	5.78	470	572	22
Crop Value per Hectare	0.35	82	311	5.55	3.06	455	401	32
Draft Oxen per Farm	0.06	83	76	5.94	9.04	493	687	23
Feasting Pots per Household	0.01	87	34	7.24	27.89	630	948	32
Percent Land in White Teff	0.22	83	113	5.87	4.36	487	493	20
Percent Land in Chickpeas	0.35	83	130	5.86	3.09	486	402	28
Yield of White Teff	0.24	69	116	2.73	4.09	188	474	108
Yield of Local Wheat	0.57	46	152	2.37	1.95	109	296	215