SOME LESSONS LEARNED FROM CONDUCTING FEDERALLY SPONSORED SURVEYS

Eugene P. Ericksen, Institute for Survey Research, Temple University

1. The Situation

The job of responding to federal contracts for statistical surveys is fraught with ambiguity and frustration. This is because there is no clear standard for the quality of data and one has to play a guessing game about which standards will be used in judging a proposal or final report. Will they be standards of data quality, standards of policy relevance, or is the agency simply interested in getting a study done for the cheapest possible cost? Caught between the Scylla of poor quality research done for a small budget and the Charbydis of high quality research done at a price no one can afford, the result all too often turns out to be that the quality of the research is poor and the budget is exceeded. Given the importance of research, the large amount of money actually spent, and the large number of qualified statisticians, precisely how this occurs is a topic ripe for investigation by a student of organizational processes. It is also a topic of immediate concern for statisticians, since the quality of our collective product does little good for the legitimacy of our field.

I suspect that one basic cause has to do with the multiplicity of desirable surveys, which results in a budget for each that is insufficient for proper data collection. Why the number of surveys can't be reduced, with the additional money available from this reduction transferred to improve the quality of the remainder, probably has to do with the large number of agencies who need research done. Many of these agencies have insufficient budgets to commission quality surveys, and they seem to be reluctant to pool their resources. Nevertheless, there are many situations where budgets could be sufficient for quality research, but money is not spent wisely. As statisticians, we can have little impact on how decisions are made on which topics to carry out government research. However, there are aspects of the problem where I think we could fruitfully bring our influence to bear.

I would like to suggest that we should try to make progress toward solving two knotty problems. One is the general lack of agreement on standards and the other is the lack of objective criteria for making statistical choices. These problems were made particularly clear to me as a member of the Review Committee for the ASA Project on the Assessment of Survey Practices. Faced with the problem of how to decide when a survey could be judged as having met its objectives, we found it very difficult to write down a set of criteria. How does one compare a clustered sample for which a 65 percent completion rate was obtained and for which sampling errors were properly computed, with a clustered sample for which an 85 percent completion rate was obtained and sampling errors were not computed? This judgment becomes even more difficult when other issues are taken into account. For example, we had to make value judgments about the importance of validating interviews, the extensiveness of checking for data reduction errors, the quality of interviewer training, and the assessment of measurement error.

It is likely that most statisticians would agree that quality is paramount and therefore probability sampling should be used, sampling errors should be computed, interviews validated, data reduction checked, interviewers trained well, and that some check on the reliability or validity of data should be made. Unfortunately, the budgets of most government agencies writing survey specifications are not large enough that all these things can be done, and we lack a methodology of choice among criteria. Moreover, there are at least two issues which divide statisticians on defining proper practice. One is the proper method of computing a response rate and the other is the advisability of cluster sampling.

Most survey organizations report a response rate as the completion rate, the number of eligible respondents interviewed divided by the number of eligible respondents contacted. In spite of generally declining completion rates, this method of reporting a response rate can often produce a pleasant result, legitimately in the 85 to 90 percent range or higher. Unfortunately, nonresponse is often dominated by noncoverage, i.e., eligible respondents actually in the sample who are not contacted by interviewers. I would like to argue that the one proper way of computing a response rate is to obtain an independent estimate of the size of the universe and then compare this estimate to the weighted sum of eligible respondents, where the weights are equal to the inverses of the respective probabilities of selection. The ratio of the weighted sum to the independent estimate is the "true" completion rate which takes into account not only refusals but also households or telephone numbers where no one was contacted, incomplete enumeration of sample households, willful concealment of refusals on the part of interviewers, and sampling units not covered by the survey process. This includes housing units missed in the housing unit listing process in an area sample and housing units without telephones in a telephone survey.

The CPS appears to be one sample survey where this comparison is consistently done, and weights are computed to adjust for differential rates of nonresponse by various demographic subgroups. There appears to be no other survey organization which consistently makes this comparison and the typical method of reporting completion rates is to use the number of eligible respondents contacted as the denominator. Emphasis on this ratio encourages fudging, because an eligible respondent who is missed by an interviewer does not count the same as one who refuses to be interviewed. Emphasis on this ratio also favors the use of quota sampling and random digit dialing telephone surveys because of the lack of concern for those who are missed by the survey process altogether. I suspect that one of the reasons the use of this procedure is continued is that it makes survey organizations look better and therefore increases their competitiveness. Estimates of total noncoverage are often embarrassingly high, and omitting such estimates significantly reduces the amount of explaining necessary to give to granting agencies. If granting organizations specified the size of the universe under study and insisted that this estimate be computed, the controversy over the proper computation of response rates could be ended.

In my opinion, there is a second area of more legitimate controversy. This concerns the ascendancy of cluster sampling and attempting to cover the entire population versus simple random sampling and not attempting to cover the entire population. On the one hand, it is typically impossible cost-wise to cover the entire household population of the United States without using some form of cluster sampling. Unfortunately, statisticians are increasingly using modern forms of multivariate analysis including log-linear modeling and logistic regression for which the error structure is not known when cluster sampling is used. Thus, some argue, it is impossible to make suitable inferences to the universe under study when we don't know how to compute sampling errors. Continuing their point, it is better to use a survey procedure such as random digit dialing or a mail survey where simple random sampling is possible, even though we know that part of the population is not being covered. Then proper statistical inferences can be made concerning the population that is covered and more speculative inferences can be made for the remainder. Given this hard choice, the added difficulty of choosing among features which all statisticians value makes the selection of a contractor from a set of competitive bids all the more difficult.

2. Organizational Factors Which Make the Problem Worse

These disagreements among statisticians weaken the basis on which rational decisions can be made by government agencies trying to decide on which survey organization to award a contract to. This decision-making process is weakened even further by two additional complications: (1) sampling theory is lacking which would aid in the choice among plans emphasizing different features of high quality research, and (2) choices about which features are most important to emphasize are not made by the government agency, either before or after the contract is awarded. Budget criteria make the final decision, and the result is that the completed research often has many unattractive features. Moreover, when the government organization isn't sure what it wants, prospective bidders are left to play a guessing game. I suspect that this ambivalence could be lessened by the more active participation of survey statisticians in the drawing up and writing of specifications for a proposed study.

Statistical procedures such as optimal allocation make it possible to balance a given reduction in variance against the corresponding increase in cost and to obtain a minimal variance sampling plan for a fixed cost or a minimal cost plan for a fixed variance. Unfortunately, sampling variation can be dominated by other sources of survey error due to unreliable or invalid measurement, noncoverage of important demographic subgroups, or sloppy data reduction procedures. We have no objective procedures for deciding on the optimal number of callbacks, or for estimating the number of questions needed to reduce measurement error for an important concept that is difficult to measure on a questionnaire. We cannot place dollar values on the personal training of interviewers relative to training by phone or through the mail. Similarly, we cannot place a dollar value on the validation of interviews. Given the disproportionate advances in sampling theory in the direction of estimating sampling errors, we lack objective criteria for assessing other trade-offs. For example, how does one compare a plan by which extra callbacks increase the completion rate by 5 percent, personal training reduces the unreliability of measurement by 10 percent, the validation of interviews weeds out the 3 percent of interviewers who cheat, and more careful editing procedures improve the reliability of measurement by 5 percent, against a plan which does none of these things but which uses optimal allocation to reduce variance by 10 percent for the same cost. These comparisons are not easy to make, even for an experienced, sophisticated statistician. Beyond measures taken to improve the bidding process, a priority area for statistical research would be to improve the methodology for assessing these tradeoffs.

In the meantime, hard choices must usually be made, and it appears that the choices are made all too often by administrators or financial officers who don't have the experience or know-how to properly confront these choices. Worse, the choices are usually not confronted until prospective contractors have submitted bids, which makes it extremely difficult for bidders to submit responsive proposals.

3. Suggestions for Improving the Bidding Process

I would like to suggest that three steps could be taken by government agencies to improve the process by which proposals are requested and selected for statistical surveys. These are (1) to make greater use of statisticians in drawing up and writing specifications, (2) to confront some of the difficult choices on survey specifications in advance and to indicate which choices have already been made and which choices they would still like to hear arguments on, and (3) to make greater use of statisticians to evaluate the collection and analysis of data after the project has been completed.

Most requests for proposals that we receive at the Institute for Survey Research give no indication about whether sampling errors should be computed, whether or not the granting agency is willing to pay for the validation of interviews and the personal training of interviewers, whether it is willing to pay for repeated measurements to evaluate the reliability of questionnaire items, whether or not substitutions should be permitted, or what kind of coverage rate is desired. A preference for probability sampling is usually assumed, and a specified response rate is sometimes given. Many of these choices could be made before the proposal specifications are written.

The present situation puts prospective contractors in a bind. Because of the standards we would like to set for ourselves, we prefer to compute sampling errors, to train interviewers in person, validate the majority of our interviews, use rigorous checking procedures in data reduction, and to collect repeated measurements to assess the reliability of our data. In fact, we insist on many of these features in our proposals, often with a religious fervor as "keepers of proper statistical practices." We have sadly lost many contracts to cheaper bidders because of this insistence on standards. The situation which often results is that the government agency is most willing to compromise on the computation of sampling errors or the assessment of measurement error. This is even more true when we subcontract for the collection of survey data to an organization which will take responsibility for analysis. Because it costs money to compute sampling errors, and because they, along with estimates of the extent of measurement error, make it more complicated to analyze data, we are often told not to compute sampling errors and assess measurement error. As a business in a highly competitive industry, we cannot afford to turn work away which fails to meet our "moral" standards, yet we are partially culpable for the poor statistical quality of some of the results. Because we find that we would confuse our interviewers and coders by relaxing our vigilance with respect to validation, training, editing interviews, and checking the accuracy of coding, the part of the survey process where we save money is in the assessment of sampling and measurement errors.

If statisticians were more intimately involved in the drawing up of survey specifications, it is likely that the hard choices would be faced in advance, and that the results of these choices could be included in the "Request for Proposals." If the specifications were rigorous, this would limit the set of competing organizations to those with the expertise to deliver the product. If the specifications were indicated to be less rigorous, organizations emphasizing high standards of research could choose not to bid. It would also be very helpful if a group of government statisticians, perhaps under the auspices of the Office of Management and Budget, got together to draw up a set of critical choices for survey specifications. Then, each RFP would have to state in advance its position on these choices, based on the amount of money available, the sample size necessary to provide useful information, and the minimum quality of information essential for intelligent decisions. The RFP would state whether sampling errors were desired, what the minimal coverage rate would be, whether or not interviews should be validated, and what type of interviewer training was necessary. It is likely that the forced confrontation of these choices would induce government agencies to opt for higher standards in order to justify the expenditure of money. This would strengthen the positions of contracting organizations and government researchers who emphasize high quality research and would likely improve the quality of research being done. If each RFP had to include a statement concerning whether or not sampling errors should be computed, it is likely that most proposals would include provisions for computing them.

Unfortunately, we know that survey statisticians and researchers in the government agencies do not have the final say concerning the choice of a survey organization. We at ISR have recently been in a situation where the research branch of an agency selected us to be the contractor, but the final decision was held in abeyance until the financial office had reviewed our budget and those of competing bidders to decide whether ours was truly cost efficient. How this was done in the absence of statisticians using statistical criteria is beyond me.

As a further check on practices, I suggest that funds should be put aside for the objective statistical evaluation of a study once it has been done. This evaluation would be public information, and would make it possible for the individuals and organizations doing the research to develop a "track record" which could be public information. For a survey with a total budget of several hundred thousand dollars, the cost of this evaluation would be a fraction of total costs. These reports would permit government organizations to check the past records of bidders.

It must be realized, however, that these procedures are likely to increase survey costs. As a result, if the standards of surveys are to be raised, a likely result is that fewer surveys would in fact be done. This could put some survey organizations out of business and result in a smaller volume of information available to government agencies. However, the quality of data would be higher and hopefully this would facilitate the decision-making process. I would argue that it is better to know you have a smaller amount of accurate information on an issue on which a decision is to be made, than to erroneously believe you have a large amount of accurate information.