DISCUSSION

William L. Nicholls II, U.S. Bureau of the Census

The papers in this session address the central issues of telephone interviewing: (1) noncoverage of households without telephones; (2) noncooperation and other forms of nonresponse when the initial contact is by telephone; (3) methods of questioning, especially modification of personal visit items for phone interviews; and (4) interviewer contributions to survey error.

The papers share several characteristics.[1/] Each considers practical procedures to improve estimates from telephone surveys or to make them more cost-effective. Where empirical results are presented, they are based on national samples of nontrivial size and reasonably well controlled study designs although several lapses from ideal design can be found. And each paper follows in the tradition of similar papers by the same authors or their organizational colleagues. Perhaps the last characteristic explains why there are no breakthroughs and only an occasional surprising result. There is, however, steady progress in most areas when viewed in the context of prior work.

Massey, Barker, and Hsiung provide the first illustration of progress. They report both marginally higher response rates and greater callback effectiveness than previously described by Fitti (1979) for an earlier but apparently overlapping time period of the same survey. The National Center for Health Statistics (NCHS) staff appears to be chipping away effectively at the obstinate nonresponse problems of random-digit-dialing telephone interviewing. Massey and his colleagues also should be commended for continued progress in the definition and measurement of telephone cooperation rates, response rates, and the percent surveyed. While use of the term "response rate" for a measure which places all repeatedly tried "ring no answers" in the denominator seems highly conservative, broader adoption of the NCHS terminology and measures would bring needed clarity to discussions of these important topics.

Three minor criticisms may be offered of this generally excellent and useful paper. First, the results on callback effectiveness would be more informative if the nature of the followup activities was made more explicit. A refusal conversion effort may consist of no more than one additional attempt to reach the household, or it may require sufficient calls to contact a potential respondent at least one more time. Without knowledge of the effort expended in followup (and the stopping rules both for initial and callback activities), it is difficult to assess either the potential effectiveness or cost-effectiveness of callbacks.

Second, it is unfortunate that callback efforts were attempted only for two-thirds of the initially unresolved cases. While field staff, as in this survey, often make intuitive judgments about cases worth and not worth additional callbacks, opportunities were missed to test the validity of these judgments and to assess the effects of thorough followup procedures.

Third, the authors surprised this discussant with their apparently serious attempt to estimate the characteristics of nonrespondents from those of initial resistors, as proposed by O'Neil (1979). Even without supporting data, it seems apparent that: (1) initial resistors may or may not resemble nonrespondents; and (2) the method is unlikely to provide trustworthy estimates of the magnitude of bias even when it correctly identifies its direction. The O'Neil strategy seems at best a method of last resort when no more reliable procedure is available. Since a better method of estimating nonrespondent characteristics was previously presented by Massey, Barker, and Moss (1979), the analysis of initial resistors seems superfluous except as a welcome demonstration of its limitations.

Peter Miller's comparison of one-step and two-step (or unfolding) satisfaction scales demonstrates both progress and unresolved problems in adapting show card items for telephone use. His procedures and results become especially intriguing when contrasted with those of his predecessors, Groves and Kahn (1979).

Miller made two changes in the original unfolding satisfaction items developed by Groves and Kahn. First, he included a second step for all three initial responses of "satisfied," "dissatisfied," and "in between." Groves and Kahn omitted a probe for the last or neutral category and observed a disproportionate number of replies falling in it. Second, Miller employed more systematic category labels throughout, such as "completely, mostly, or somewhat satisfied" for those initially answering "satisfied." Groves and Kahn used more idiosyncratic (but vivid) categories, such as "good," "bad," and "mixed" at the first level and "delighted," "pleased," and "mostly satisfied" for those answering "good." It is less clear that this change was an advance.

Two results of Miller's paper seem to suggest weaknesses of the two-step or unfolding method. First, when compared with one-step versions of the same items, the two-step versions produced somewhat higher mean satisfaction scores and occasional heaping in the "completely satisfied" category. However, this result may reflect the polarity of the questions rather than their number of steps. The one-step question is defined for respondents from the dissatisfied pole by the instruction that "One stands for completely dissatisfied and seven for completely satisfied." The two-step item reverses the polarity by first asking: "Would you say you are satisfied dissatisfied, or somewhere in the middle?" Locander and Burton (1976) initially proposed the unfolding method to minimize bias resulting from the common respondent tendency to choose first mentioned categories, but in Miller's unfolding question respondents choosing the first mentioned categories at both levels are led into the "completely satisfied" response. This may explain both the higher satisfaction scores of the two-step method and the special heaping in this category.

The second result suggesting a weakness of the unfolding technique is the smaller average inter-item correlation for two-step than one-step versions of the same items. Groves and Kahn report exactly the opposite result; in their earlier study, the unfolding versions had the higher inter-item associations. Since the two studies differed in the specific satisfaction items examined, the measure of association employed, and possibly other relevant ways, no firm conclusions may be drawn, but the possibility exists that the more vivid category labels of the Groves and Kahn unfolding questions strengthened their reliability.

The reviewer fully concurs with Miller's call for additional systematic research to identify effective approaches to scale measurement in telephone interviewing. The list of variables requiring attention may be even longer than Miller implies. They include: (1) the number of steps; (2) the necessity of probing each initial unfolding category; (3) the polarity of each set of items; and (4) the verbal labels chosen at each level.

Groves, Magilavy, and Mathiowetz present an intricate and provocative analysis of monitored interviewer behavior and interviewer variability. Both the topics and methods of analysis parallel those of previous investigations by the same authors and their Michigan colleagues, but this paper analyzes a larger sample, examines different survey items, and attempts to account for interviewer variability by inappropriate interviewer behavior observed in monitoring.

The authors express apparent surprise at two of their major results. First, the values of p^*_{int}, the magnitude of interviewer variation, for the 15 health items examined here are smaller than those found in previous analyses by Groves and Kahn (1979) and Groves and Magilavy (1980). In the present analysis, interviewer variability approaches trivial size. The contributions of telephone interviewers to survey error, previously assumed to be a pervasive, major problem, suddenly appear in this survey to be a nonproblem. This unanticipated success cries out for explanation.

Modesty almost forbids the authors from considering one plausible explanation. The Michigan group has devoted years of effort to the reduction of survey error through careful interview design and usually thorough interviewer training and supervision. Could they have succeeded here even beyond their own expectations and without one clearly demonstrable breakthrough to account for their success? This possibility cannot be ruled out; interviewer variability may prove reducible to trivial limits if one works at it hard enough. But a string of successes, rather than just one, would be necessary to support this optimistic interpretation.

An alternative explanation offered by the authors is that the 15 health items examined here are less susceptible to interviewer effects than the attitudinal items analyzed previously. A review of previous studies conducted or summarized by the authors suggests that factual items may demonstrate less interviewer variability than attitude items, but the p^*_{int} values reported here are small even in comparison with the factual items of previous studies. We will have to await further results before the apparently anomalous findings of the present study are more clearly interpretable.

The second major result which Groves, Magilavy, and Mathiowetz find surprising is the apparent lack of relationship between interviewer variability on an item and inappropriate interviewer performance monitored on that item. Perhaps too little interviewer variability remains to be explained; no relationship is possible because there is no meaningful variation in the dependent variable. However, even if the magnitude of interviewer variability had been larger, it is doubtful that the study design could have demonstrated such a relationship.

The measures of interviewer variability, p^*_{int}, are based on an average of about 58 cases per interviewer, but for reasons of cost not all interviews were monitored. An interviewer's reading of a specific item typically was observed only 5 to 10 times, although inappropriate behavior typically occurred only for about 1 reading in 9. Clearly insufficient observations were made to obtain reliable measures of interviewer performance by item. In the scatterplots presented by Groves and his colleagues, the values of the horizontal axes are apparently closer to random variables than estimates of individual interviewer performance during the course of the field work. A revised study design, which permits reliable estimates of monitored behavior throughout the field work period, seems required before hypotheses relating interviewer variability and interviewer behavior can be meaningfully tested.

The paper by Casady, Snowden, and Sirken continues development of dual frame (household and telephone) sampling designs introduced by Casady and Sirken (1980) at last year's meetings. Since the dual frame strategy for concurrent personal and telephone surveys, proposed and theoretically grounded by these investigators, provides the means of avoiding the undercoverage bias of telephone surveys while realizing their cost advantages, it would be difficult to find a more practical, significant, and exciting development in survey sampling at this time. The present paper continues the important tasks of laying out specific designs, refining cost estimates, and educating a wider (and less mathematically sophisticated) audience on work in progress. The last is especially appreciated and should receive even greater emphasis in future presentations.

The paper by Burke, Morganstein, and Schwartz attempts to move beyond the landmark article on random digit dialing by Waksberg (1978) in two ways: (1) through a more detailed explication of field costs and (2) by development of an optimization model which includes a term for nonresponse bias. The paper apparently represents work at a very early stage of development. Central terms, such as "response rate" and "nonresponse bias" are undefined; the paper consists largely of background material and future goals; and mathematical development is promised but not completed.

The effort to clarify cost elements of multiple stage, random-digit dialing is welcome, but the authors appear caught between realistic assessments of cost components and simplifying assumptions necessary for mathematical solution of their cost model. They recognize that the cost of a telephone call varies with the stage of sampling and screening, but to facilitate mathematical development they assume that all calls cost the same. The benefits of this assumption, or a reconciliation of the model with the real world, remain to be demonstrated.

The attempt to construct an optimization model incorporating an exponential function of the number of callbacks, serving as a proxy for nonresponse bias, strikes this discussant as a dubious line of development for several reasons. First, it ignores other important forms of bias and error, including noncoverage and interviewer variability. Second, it assumes that nonresponse bias is a simple function of the response rate. The paper by Massey and his colleagues well illustrates the complex and unpredictable ways in which nonresponse bias for an estimate varies with the study response rate. Third, the development assumes that the nonresponse rate is a simple exponential function of the number of callbacks. Illustrations presented by the authors support the view that the number of callbacks is one (and probably the major) factor influencing the response rate, but the sponsorship of the survey, its content, and perhaps the nature and timing of calls also may play major roles. A model concentrating only on mechanical completion of specified numbers of calls could mislead rather than inform.

In view of the early stage of this paper's development, the discussant recognizes that these criticisms may be premature. The extreme assumptions made by the authors could conceivably lead to instructive results. Since the utility of their work remains to be demonstrated, one can only wait to learn if that utility justifies the assumption made.

In summary, the authors of the five papers are to be commended for their continued and provocative efforts to push back the frontiers of surveys by telephone. Their next developments should be eagerly awaited.

FOOTNOTES

[1] The discussion is based on the papers as presented or as received prior to presentation.

REFERENCES

Casady, Robert J. and Monroe G. Sirken. "A 1980 Multiplicity Estimator for Multiple Frame Sampling," Proceedings of the American Statistical Association, Section on Survey Methods.

Fitti, Joseph E. "Some Results from the 1979 Telephone Health Interview System," Proceedings of the American Statistical Association, Section on Survey Methods.

Groves, Robert M. and Robert L. Kahn. Surveys 1979 by Telephone, New York: Academic Press.

Groves, Robert M. and Lou J. Magilavy. 1980 "Estimates of Interviewer Variance in Telephone Surveys," Proceedings of the American Statistical Association, Section on Survey Methods.

Locander, William B. and John P. Burton. "The 1976 Effect of Question Form on Gathering Income Data by Telephone," Journal of Marketing Research, 33, 189-92.

Massey, James T., Peggy R. Barker, and 1979 Abigail J. Moss. "Comparative Results of Face-to-Face and Telephone Interviews in a Survey on Cigarette Smoking," presented at the American Public Health Association Meetings, November, 1979, New York.

O'Neil, Michael J. "Estimating the Nonresponse 1979 Bias Due to Refusals in Telephone Surveys," Public Opinion Quarterly, 43:2, 218-32.

Waksberg, Joseph. "Sampling Methods for Random 1978 Digit Dialing." Journal of the American Statistical Association, 73:361, 40-46.