

DISCUSSION

Mary Grace Kovar
National Center for Health Statistics

These papers, which may represent the best work being done on measurement and response problems in survey research, offer a stern reminder of how little carefully designed research is taking place in this area.

In the course of this meeting there have been many papers in which the authors develop finer and finer points in the theory and mathematics of data analysis. We are now past the point where one starts with the statement "Assume the covariance is zero". We can compute the full variance - covariance matrix for complex surveys and modeling is now possible for the data.

It is noteworthy that the authors of papers on the mathematics of data analysis do not address the issues of whether the data they're analyzing are really measuring what they purport to be measuring. It is also noteworthy that, for the most part, the authors of papers on measurement and response problems are not using the more sophisticated methods of analyzing the findings. This apparent lack of communication is, I think, unfortunate for the field of survey research.

Sophisticated analysis is pointless if the data have systematic biases due to questionnaire design, inability of respondents to report accurately, survey design which fails to include the entire target population, or the host of other sources of bias in surveys. Conversely, the study of measurement and response problems should take advantage of the analytic procedures which are now available.

The papers presented on this session can be divided into groups.

There are two papers reporting on the validity and reliability of sensitive topics - alcohol use and illegal activities - and one that is on overview of response errors for sensitive topics. There is one paper reporting some of the results of a survey experiment. There are two papers that use statistical analysis to estimate and compensate for errors in the data.

I have four comments on the papers that address validity, reliability, and response bias of so-called "sensitive" topics. The first is that the field of survey research needs to define validity and reliability, to lay out sufficient means of measuring both, and to use good statistical practice in setting levels of acceptance and rejection. While it is always necessary to look at levels of nonresponse and to make comparisons with other questions on the same questionnaire, it is not sufficient to use those measures alone to state that the data are valid or reliable.

The American Psychological Association has published standards for validity and reliability. The American Statistical Association needs to do the same.

The second point is that sensitivity should be based on the perceptions of the target population not on those of the survey designer. I wonder who decided that the topics addressed in these three papers presented here were "sensitive"? I suspect that the decision was made by white, middle-class males some 20-25 years ago. Their perceptions of what was sensitive for the entire population may not have been valid then and are less likely to be so now. We need to look at variation among population groups because what's sensitive to one group may not be sensitive to another, and we need to be aware of the possibility of change over time because a topic that was sensitive in 1960 may not be sensitive in 1980.

The third point is that, as the paper by Marquis et al mentions, the problem may not be sensitivity at all. It may be poor questionnaire design. Statisticians are notoriously conservative and are likely to cling to questions because that's the way "it's always been done" or "that's the way someone else did it". If they do, poor questionnaire design will be perpetuated as poor response to sensitive questions.

The fourth point is that every good study should include comparison with other data for as many of the estimates as possible, not just the sensitive ones, and should make provision for studying over-reporting as well as under-reporting.

This leads to the paper on the survey experiment - one of the rare cases of an experiment designed to test survey procedures. I have a feeling that too many design factors were built in: self vs proxy respondents, personal vs telephone interviewing, and several models of telephone interviewing. All those factors may have overwhelmed the ability to look at some basic issues such as bias introduced by using proxy respondents which is the subject of this paper.

The striking finding of this research, if it is upheld, is that proxy respondents on the telephone survey do not report lower levels of health events although proxy respondents on personal interview surveys usually do. The review of studies in the Marquis paper revealed that the only under-reporting bias was for chronic conditions - a health event. That's based on old research. Perhaps there has been a change over time, perhaps the issue is not self

vs proxy respondents but a "knowledgeable" respondent, and perhaps the real problem was poor questionnaire design in the personal interview survey.

Two additional comments on this paper are: first, the use of a single respondent for the household does not necessarily decrease variance because of the increased sample size. The design effects for the NHIS are not negligible and there is evidence that included in the household clustering is a measurement problem: the respondent may be reporting that other family members are like her whether they are or not.

Second, the Verbrugge hypotheses that male-female differences in utilization are due to self-proxy respondent differences is not upheld by other data such as the National Ambulatory Medical Care Survey and the National Hospital Discharge Survey. Utilization measures from the National Health Interview Statistics and from these two surveys reveal the same differences even though the latter are based on reporting by medical care providers rather than household respondents. [2]

Finally, there are the two analytic papers.

I was delighted to read the simulation paper which is, as far as I know, the only

example of its kind. This paper does relate the problems of measurement to the analysis and the inferences that can be made from the data. The authors demonstrate that response errors can lead to biased estimates and invalid inferences. We need much more research on the topic before we know how often that happens.

The Sabertehraniz paper used canonical analysis to partition the variability of rating scale responses into the components due to characteristics of the scale, type of measurement instrument, data collection mode, and the environment in which the data were collected. He too found that such factors could distort the research findings. The paper would have been much stronger, however, if the author had used less jargon and had presented his rating scales and research results.

References

1/ Standards for Education & Psychological Tests. Published by the American Psychological Association, Inc., Washington, D.C. 1974.

2/ Kovar, M.G., The Elderly Population: Use of Medical Care Services by Men and Women in their Middle and Later Years. Presented at the Annual American Public Health Association, Detroit, Michigan October 1980.