DISCUSSION

Patrick J. Cantwell, Bureau of the Census
Statistical Research Division, Washington, DC 20233

Mathiowetz and Lair

In this paper, functional limitation in the elderly is measured by the number of activities of daily living (ADLs) one can easily manage, as reported in the National Medical Expenditure Survey (NMES). The reader can see how respondent interpretation makes it difficult to count the actual number of such items. Mathiowetz and Lair describe several problems: the confounding of the target characteristic with the propensity to be a respondent, the tendency for proxy respondents to report more functional limitations, the misinterpretation of the survey questions by respondents, the group of 125 new interviewers in the fourth round of the survey, as well as other common sources, such as conditioning and response variability. As the authors indicate, the analysis can account for certain difficulties, such as the new interviewers.

If there is a problem counting ADLs in a given round (wave) of the survey, how much more difficult is it to measure the *difference* in the number of ADLs as measured through time? Does a change of one ADL from Round 1 to Round 4 imply a real improvement or decline in the functional ability of the respondent, or might this small change be due to measurement error?

Further impairing reliable measure of change is the exclusion of most of the sample respondents (85%), those who had no functional limitations in either round of the survey. If we add to this group those who reported at most one more or one fewer ADL in Round 4, we exclude over 95% of the sample. That is, fewer than 5% of the respondents indicated a change of two or more ADLs. Even if we eliminate those without functional limitations in either round, about 70% of the remaining respondents report a change of one ADL or fewer.

The authors present two multivariate logistic regression models based on the NMES data. One model predicts the probability of functional improvement from Round 1 to Round 4, while the second predicts the probability of functional decline. The two models consider a multitude of variables including health factors of the respondent, demographic characteristics of the respondent and interviewer, and administrative details of the interview.

As insightful as the models are, the first--for predicting the probability of improvement--suffers from a restriction of the sample. Because only respondents who reported functional limitations in Round 1 can improve (as measured by the number of ADLs they can handle), only they are used in determining the model. This eliminates almost 90% of the sample. For predicting decline, only the few percent who reported limitations in *all* ADLs in Round 1 are removed from the modeling process.

This vast difference in sample size for the two models may explain why so few variables are significant contributors for predicting improvement compared to the number for predicting decline. Further, we have less confidence in the predicted probabilities of improvement. The authors demonstrate a surprising example where the respondent in the given base case shows a 75.3% chance of improvement under the model. By changing only the interviewing experience and education *of the interviewer*, the predicted chance of improvement drops to 33.3%.

Finally, a main premise of the authors is that data from the NMES imply an unexpectedly large amount of improvement in functional status among the elderly. Part of this is properly attributed to response error. Perhaps another part can be explained by the 226 individuals who died or were institutionalized between Rounds 1 and 4. Had they been included, it is likely that the model for predicting improvement in the elderly would have produced less optimistic results.

Berlin et alii

There is a substantial body of literature on the effect of monetary incentives on response rates in surveys. The paper by Berlin and her co-authors goes further by studying this effect on response rates, the total cost of data collection for the survey, and respondent performance.

For this particular survey, the National Adult Literacy Survey (NALS), nonresponse can be an especially serious problem. Key characteristics affecting literacy performance--such as education and income levels--are often highly correlated with response in surveys of this type. A differential nonresponse rate can introduce a significant bias in the estimates. The authors emphasize how important incentives can be in certain demographic subgroups to prevent a serious selection bias.

In the paper we find an interesting and detailed study of the effect of incentives broken down by various groups. Particularly significant are differences by race and ethnicity. It would be interesting to explore whether these characteristics are merely proxies for income level in predicting response. (Perhaps income level is not available for respondents.)

In the results, it is interesting that, although response is significantly better as the incentive increases from $0 to $20, response is only marginally higher as the incentive rises from $20 to $35. We wonder what the response curve looks like as a function of the incentive. Where does it begin to flatten out? Further, what is the optimal incentive value when considering the total cost of the survey? The authors provide a brief yet illuminating analysis of the average cost per completed interview. We see that an incentive of $20 can save about $31 on average compared to no incentive. An incentive of $35, however, saves only slightly more ($36) and is not cost effective.

The authors might continue this line of research by considering work done by R. Bolstein or others. Addressing response in mail surveys, Bolstein has studied the use of different combinations of incentives and follow-up mailings. His work has revealed combinations which have comparable effects on response. Perhaps the authors of this paper could investigate which combinations of incentives and follow-up yield similar results in terms of respondent performance and total cost for the survey.

Thomas and Dingbaum

Reinterview is the topic of this paper and the next two, although each addresses a different issue. The Thomas and Dingbaum paper describes the Content Reinterview Survey (CRS) for the 1990 Census--its procedures and an initial look at the error measures for several items on the census.

When reviewing the success of the CRS, it is meaningful to weigh its differences with reinterview for the Census Bureau's major demographic surveys. The fundamental differences are in purpose and procedures. The ongoing surveys use reinterview mainly to control falsification and to monitor the performance of the field interviewers. A supervisory staff person in the field usually reinterviews one person in the household shortly after the original contact. However, the census, being a one-time operation, is mainly concerned with the quality of its data, that is, the extent of response error. In the CRS, CATI was used to reach everyone in the household 15 years of age or older (15+) sometime between September and December, 1990.

Further, while most ongoing surveys prescribe reinterview procedures to measure either response variance or response bias, the CRS addressed both of these goals. For some characteristics, such as Spanish origin, the CRS measured response variance; for race and others items, the CRS measured response bias.

When trying to gauge response variance, one typically tries to replicate the original interview and context. There are several key aspects of the CRS which differed from the census, most of which the authors note:

- Mode of interview. Mailout or personal enumeration in the census, telephone in the CRS.

- Time (recall). The CRS was conducted from four to nine months after the census. This may be a problem only for items where the characteristic can change (employment, rent or mortgage, etc.).

- Conditioning. The responses in the CRS could be affected due to answering the census the previous spring.

- Different conditions. The overall context--social or economic--may have changed between contacts.

- Respondent(s). While the census form may well have been completed by one household member, the CRS tried to interview all individuals 15+.

When evaluating measures of response bias, we generally look for areas where the reinterview improves on the original interview. As we mentioned, the CRS attempted self-response from all individuals 15+. Further, a series of probing questions were used in the CRS, avoiding the more confrontational approach sometimes used in reconciled reinterviews.

Sinclair and Gastwirth

When measuring response errors through reinterview, survey analysts like to make several assumptions. As the authors indicate, first we assume that the error rates in the original interview are conditionally independent of those in the reinterview. Although this is sometimes questionable, it is of less concern for the Content Reinterview Survey (CRS), conducted four to nine months after the census. When measuring response bias, the error rates in reinterview are usually assumed to be very close to 0. Even with specially designed instruments for the recontact, this is unlikely. When measuring response variance by replicating the original interview, we usually assume the error rates in reinterview are about the same as in the original. Often this is also a specious supposition.

In this paper, Sinclair and Gastwirth try to measure response error without relying on the validity of the above assumptions. To start, they investigate the usual measures for response bias and variance--the net difference rate (NDR) and the simple response variance (SRV). They show that, if the assumptions do not hold, the usual estimates for π (the prevalence rate of the characteristic being estimated), NDR, and SRV are biased. Further, the bias in NDR and SRV are functions of π and the error rates of the interviews.

By using a method of Hui and Walter, the authors hope to forego the usual assumptions in favor of a different set. As before, it is assumed that the error rates for the two interviews are independent. Now we select two or more subpopulations such that the *interview error rates are equal across subpopulations*, but the prevalence rates are unequal. Although I cannot say that the usual assumptions are easier to satisfy, I still have concerns about the validity of this one. Here, we must determine subpopulations where the prevalence rates differ, but then assume that the error rates are the same in each subpopulation.

The Hui and Walter procedure is applied to two items from the CRS, Spanish origin and employment. This selection allows us to see how the method works with small and large values of π. Males and females were selected as the subpopulations, having different prevalence rates for these characteristics. While the authors assume that the interview error rates are the same for males and females, they state that procedures used "do not allow for the verification of this assumption."

The analysis concludes with an insightful look at deviations from the Hui and Walter assumptions. By allowing the error rates to vary between the two populations, one can see how robust the new procedure is. My only suggestion to the authors would be to extend the analysis with larger values of the error rates. In the analysis, the largest error rates considered for the original and reinterview are .065 and .026, respectively, perhaps unrealistically low.

Brick and West

I chose to end with this paper because I feel it demonstrates the future of reinterview for most of our demographic surveys. Although the authors do a commendable job describing the measurement of response reliability in the 1991 National Household Education Survey (NHES:91), I find the mode of reinterview used to be the most interesting and promising aspect of the paper. In most reinterview settings, one tries to measure response variance *or* bias--not both, because the optimal conditions are different for each. In fact, until recently, the Census Bureau conducted a split reinterview sample for the Current Population Survey. Some reinterviews were done without reconciliation to measure response variance; others were done with reconciliation to measure response bias.

The Thomas and Dingbaum paper recounted how the Content Reinterview Survey for the census measured response variance on some items and response bias on others. Brick and West describe a CATI reinterview procedure where we can measure response variance *and* bias *on each characteristic.*

The reinterview is conducted in two segments. First, to estimate response variance, the conditions are kept as close as possible to the those of the original interview. These include using the same CATI system, respondent, and question wording, even interviewers with the same level of experience. Just as significant, the CATI system ensures that the interviewer cannot yet see the original responses.

When this part of the reinterview is completed, the responses are locked in, the original and reinterview responses compared by the computer, and reconciliation screens prepared for the second segment. The same items can then be measured for response bias.

Perfect? Unfortunately not. As the authors point out, if the original interview could be replicated exactly, we would expect the proportions of errors attributable to the original interview and to the reinterview to be about the same. However, more than three times as many errors in the NHES:91 were associated with the original interview.

Several reasons are offered by the authors. Knowing reconciliation will follow might cause the reinterviewer to be more careful. This possibility may obstruct our attempt to measure response variance. Another explanation is *internal consistency*--the tendency for respondents during reconciliation to defend their latest answer. This might hamper our effort to measure response bias. A third possibility: could it be that a reinterview workload involves fewer cases, allowing the reinterviewer to expend extra time and attention? (This is not true for many Census Bureau reinterview workloads.) During the session, a member of the audience suggested another reason. The original question may spark an interest in the respondent, causing him or her to check the correct answer between the time of the original interview and the reinterview.