DISCUSSION

Gary M. Shapiro, U.S. Bureau of the Census

This is an important, significant session. Arguably, "Data Quality" is the most important topic regarding sample surveys which is being addressed at this convention. To the best of my knowledge, this is the first time that a convention session has been titled "Data Quality". I hope this is an indication that there is now more work and emphasis on this topic than there has been in the past, and would hope that all future ASA conventions will be able to have data quality sessions.

The papers in the session are diverse, covering different aspects of quality. I found all the papers very interesting and valuable, and encourage the authors to continue their work in the same areas and present more papers dealing with data quality for future conventions.

Four of the five session papers was all I felt I would have time to discuss. Thus, I decided arbitrarily and capriciously not to discuss the census undercount paper by Mulry-Liggan and Hogan. This is not a reflection on the paper, which I think is excellent. I will now discuss the other four papers in turn.

I. MILLER ET AL PAPER

The topic of this paper, comparing self response to proxy response, is an important one. There have not been many efforts to assess the accuracy of proxy responses and more are needed. Previous work has been mostly addressed to response bias, so this paper is particularly welcome because it examines response variance.

The experiment described in the paper was very well designed. By first interviewing a random respondent (RA) and then a most knowledgeable respondent (MK) the authors have given themselves a good solid basis for comparing self and proxy respondents. Let me emphasize one particularly important point the authors make - male vs. female comparisons are potentially quite biased because proxy responses are given for males much more frequently than for females. My specific comments on the paper are in two general areas: Methodology and analytical conclusions.

My first question on methodology is whether the methodology might tend to lead to worse data for MK. I am interested in how it was explained to respondents why reports were going to be obtained from both RA and MK. It would appear possible that the MK would be a less motivated respondent than the RA. There are a number of ways that conversation between RA and MK prior to the interview with MK could negatively affect the quality of MK responses. For example, MK could be told by RA that the way to get the interviewer out of the house quickly is to answer "no" to everything.

A second methodological question has to do with the level of nonresponse. In particular, are there many housing units at which interviews were completed with RA but not with MK? High levels of nonresponse could bias the data comparisons and significantly affect conclusions. The authors left out about 1/3 of the housing units from their analysis because it was more than two days between the RA and MK interviews. This by itself might result in significant bias, and a high nonresponse rate in addition could have very serious effects.

With respect to the data analysis, one questionable analytical method was used. In testing a null hypothesis, it is improper to say that the hypothesis is true just because you are unable to reject it. The authors, however, say they will do this in the section of the paper "Models of Agreement and Disagreement in Repeated Measurements" and then proceed to do so several times. For example, they state that mothers are unbiased proxy reporters more often than are wives on health complaints. The authors, however, have only determined the number of cases where they can definitely conclude that mothers were biased reporters. In the other cases, all that can be said is that there is no evidence of biased reporting. One cannot ever firmly conclude from the analysis done that mothers are unbiased reporters. Furthermore, if in truth mothers and wives had identical levels of reporting bias, one would expect to find more significant differences for wives because there were many more RA's who were wives than mothers. Thus, I believe no valid comparisons between wives and mothers are possible.

On a related point, the authors say that when the differences were significant, the mothers had bigger differences than the wives. This is probably because of the larger sample size and the smaller standard errors for wives, rather than due to any real population differences.

I have one final comment on the analysis. In the "Summary and Conclusion", the authors say "It is at least reassuring that we fail to find dramatic effects by some important proxy and household attitudes...". The standard errors on the estimates are quite large and can easily prevent large real differences from turning out significant. Thus, I think there is a high degree of uncertainty here and I am much less reassured than are the authors.

II. TIPPETT PAPER

Any time administrative data is used to evaluate quality of survey data, it is highly worthwhile. Such evaluations are not done often enough and thus this paper is a valuable addition to the literature. Year built, which is evaluated here, is particularly important because it affects coverage in Census Bureau household surveys. In part of the country, we do area sampling for the old construction and get new construction (since the last census) by sampling new construction permits. We ask year built in area sample units to determine if a unit should be excluded from the area sample because it is covered by the new construction sample. Thus, if a unit is erroneously reported as built before the last census, it has a double chance of selection. If it is erroneously reported as built since the last census, it has no chance of selection.

I have one major question on the paper. One of the "4 A's" Mrs. Tippett discusses is the accuracy of the administrative records being used. In this regard, I wonder if the assessor is always right. Especially in the situation where the census and the American Housing Survey (AHS) agree, the assessor might be wrong. Speaking personally, I believe that in some years my estimate of my home's market value was far superior to the tax assessor's. This occurred during a period of rapidly increasing home values in which the assessment seemed to seriously lag the market.

I have one general comment. I had hoped the analysis in this paper would show AHS data quality to be better than the census since AHS has better trained interviewers and more control over the survey process. I am disappointed that there seems to be no basis for concluding that AHS is better.

My final comment is that there are three things not in the paper that I would have liked to see:

1) It would have been interesting to have more comparative analysis, particularly for the census vs. AHS, covering such questions as when is AHS quality better than census quality.

2) The paper would benefit from the inclusion of sampling errors. For example, I wanted to make comparisons between the census and AHS but did not know when comparisons were meaningful without standard errors to refer to.

3) It would be valuable to look at the magnitude of discrepancies as well as the number of discrepancies. For example, if for housing value there was a lot of discrepancies between the $50,000-$60,000 category and the $40,000-$50,000 category, this isn't too serious. Many of these discrepancies could be very marginal, e.g., $51,000 vs. $49,000. However, a number of discrepancies between the $50,000-$60,000 category and the less than $20,000 category would be extremely disturbing.

III. KOSARY ET AL PAPER

I particularly applaud the goal expressed in the introduction of the paper of designing quality assurance to not just measure quality but also to build quality in. This is clearly what we ought to be doing and represents a real improvement over normal survey practice. In reading the paper, I was not convinced that what was done is all that different from normal, good survey practice. I feel that an excellent job was done - the research was solid, the sample design is good, and the monitoring is good. But either the introduction led me to expect too much from the paper, or there are some very innovative things done that I failed to fully appreciate when I read the paper.

In addition to this general comment, I have several specific questions and comments:

1) I'm interested in knowing more about the availability of information from the edit program. How quickly is it available? Is it available to the field staff or only to Washington staff?

2) I think the process audits the Bureau of Labor Statistics (BLS) are doing are very good. I would encourage BLS to do audits in additional areas and to periodically repeat audits for the two areas already done.

3) The paper discusses problems with imaginary boundaries for segments. For Census Bureau surveys, we enlarge segments as necessary to avoid ever having imaginary boundaries. Is it possible for something similar to be done for the Consumer Price Index Housing Survey?

4) My final two comments are on the section on monitoring reports. I think one needs to be careful about having one set of national standards that gets applied uniformly. There are local differences that ought to be considered. For example, achieving a 1 percent refusal rate for a survey might be easy in a rural southern county but impossible in New York City. It is of no help to field staff in New York City to keep telling them that their refusal rate is unacceptable <u>unless</u> one can also tell them what they can do to improve it.

5) On the monitoring reports themselves, I think it is important to produce longitudinal reports and charts as advocated by Deming and Juran. Without longitudinal information, field staff can't tell if they've got a new problem or just a continuation of old problems. Also, well designed graphs are much easier to use and interpret than tables.

IV. NOVOTNY PAPER

Both this paper and the Kosary paper describe major efforts by BLS to improve the quality of their surveys. To be able to present two such papers at a single convention is quite impressive and extraordinary.

The analysis presented in this paper is exemplary, avoiding arbitrariness that's common in inspection plans by determining the effect of different decisions and arriving at decisions rationally. The plans decided upon are highly effective in meeting the objective of maintaining quality - an initial error rate of 5 percent is reduced to 0.25 percent. The plan should also be effective in achieving a second objective of identifying problem areas. I have some concern over effectiveness in achieving the third objective of providing feedback to workers.

Consider that a worker's actual error rate may be constant over time or it may change. First, take the situation of constancy. If the error rate is near 0 percent, the inspection rate will quickly go to 25 percent and will stay at that level. This is okay although 25 percent inspection is much higher than needed for very low error rates. If a worker has a "moderate" error rate, this inspection plan will usually result in a 25 percent inspection rate but sometimes a 50 percent or 100 percent rate. The changes in inspection rates will be entirely random, with workers occasionally getting the impression that their work has become unsatisfactory when their real underlying error rate is unchanged. The reason for this situation is that no historical data is used in determining inspection rates. A worker could go for a long time with no errors, but one critical or six other errors is enough to classify the work as unacceptable. What is needed, I believe, is to base decisions on more than just the last 10 forms.

The other situation which may occur is that a worker's performance has actually gotten worse. Here the inspection plan will operate fine, but so would looking at a longer historical record. I would encourage Mr. Novotny to apply the control chart ideas of Deming and Juran, looking to see if statistically significant changes in error rates occur and only then increasing the inspection rate and giving negative feedback to workers. What is presented in this paper is essentially an old-fashioned inspection plan rather than a Deming type of process control.

My final comment is addressed to BLS and is not intended in any way as a criticism of the paper. What exactly are the plans for feedback? The feedback, I believe, is what is most important. I would encourage BLS to treat the initial 5 percent error rate as unacceptably high and work hard to provide constructive feedback to workers or/and to make basic system changes to reduce errors. With significant reduction, the base inspection rate of 25 percent could be greatly decreased.