

DISCUSSION

Daniel Kasprzyk, U.S. Bureau of the Census
Washington, D.C. 20233

The papers presented in this session are for the most part concerned with errors of measurement, and, in particular, errors of response. They are papers concerned with large-scale government surveys, and they are papers prepared not just by statisticians or social scientists employed by the government, but by individuals in the private and academic sectors as well. This latter point is important. The size, scope, and implications (both fiscal and policy) of government data collections make it virtually impossible for government statisticians to carry out their mandate, requiring substantial assistance from and collaboration with the other sectors.

Except for the paper by Cohen which spends considerable time discussing frame development and the sampling from it, the papers in this session are concerned with response error: 1) can proxies who report for youth, aged 16-24, correctly classify the labor force status of the subjects? Obviously, issues of questionnaire design and respondent rules come to the fore here; 2) what is an appropriate method for obtaining expenditure data on apparel? The advantages of small-scale testing and focus groups help Cantor and his colleagues in their work; 3) can a series of questions in a RDD survey be used to correctly ascertain geographic identifiers for sampling?; 4) can a single simple question about family income be used to identify an important subgroup for oversampling?

These papers are good efforts at addressing difficult problems. The results inform us of the limitations of our questionnaires and survey procedures and, as always, point to the need for more research. The papers are good examples of the kind of review and evaluation large surveys funded by public money should undergo. In several instances, the papers seek ways to improve the efficiency (and reduce costs) of the data collection (Marker et al, Moeller and Mathiowetz). A few remarks on each of the papers follow below.

TANUR AND SHIN

This paper is the result of an excellent joint program of the National Science Foundation, the Bureau of Labor Statistics, and the American Statistical Association that facilitates the involvement of academic researchers on government statistical issues through a "fellowship" program. Previous research by the authors has suggested that youths

conceptualize the job search activity differently than adults and this can undoubtedly effect the youth unemployment rate--an important and highly political statistic.

The issues addressed in the paper are "who are the respondents?" and "how are they related to the subjects (the person the data are about?)."

What have we learned?

1. Proxy reporting is high in the CPS--about 45%, and is higher for males than females.
2. Proxy reporting decreases with increasing age among youth, but does not vary greatly by race.
3. Proxies are predominantly female and are usually parents for 16-24 year olds.
4. The unemployment rate is usually higher for self-reporting youths.

Tanur and Shin produce much information for the CPS user community about reporting quality in the CPS. These results help us to understand as well as raise questions about what the unemployment rate means. As you read the paper you wonder if other important policy subgroups have similar reporting patterns and how their reports affect the statistics we calculate; the Hispanics, for example. The authors are taking precisely the right approach in saying that the next steps will be to model the self-selection process and reporting process, and I encourage them to use economic information from the CPS March Supplement for their model.

Where will this analysis lead? Will respondent rules or procedures change when the research becomes more definitive? Will models be developed to adjust the youth unemployment rate? Will the kinds of results Tanur and Shin have shown be available in routine documentation produced for the CPS? I hope so, because the results are only helpful if the broad range of analysts who use CPS data know of them.

CANTOR, KEIL, GREENLEES AND ROSE

Cantor and his colleagues have presented a systematic and thorough description of a study aimed to improve the collection of expenditure data on apparel through the introduction of a new diary survey on apparel expenditures. The paper provides a complete and organized approach to the process of providing new instrumentation for a portion of the Consumer Expenditure Survey.

Both the sponsoring agency and the authors of the paper should be commended on their attempt to develop a "simple" collection instrument that tries to collect data on difficult concepts. The diaries tested were an attempt to motivate the respondent through the simplicity of the instrument and it appears to have worked. What have we learned?

- Simplicity in design and layout appears to be extremely important for maintaining motivation; respondents feel comfortable and are motivated to respond when the form looks uncomplicated.
- Respondents tend to use the "assists," such as the table of contents and tabs when responding.

The problems with the diary seem to be the ones expected:

- Classifying expenditures into domains.
- Classifying size of apparel.
- Items that are difficult to classify might be missed.

These are significant problems that are not about to go away. A continuing program of research is necessary. Some specific points concerning the paper:

1. Burden reduction is the reason given for trying the new approach. Another reason, not given by the authors, but certainly apparent is the perceived improvement in the quality and completeness of reporting. As far as burden reduction, the authors did not indicate how much burden would be reduced.

2. Cost played a role in some early decisions for the study; in particular, the decision on the consumer unit diary versus the individual diary. Was there an assessment of costs versus quality/completeness of response under these two designs? Or was the decision based on intuition? Using individual diaries seems more complicated, perhaps substantially so, but would the data be substantially more complete?

3. Because of cost and time constraints, staff judgement played a significant role in the organization of the diary. Based on the results, staff have excellent insight, but it does seem illogical to not have done more background research concerning the organization of the diary given the importance of this survey. Was it simply not worth spending the time on organizational principles? Was it that obvious what was needed?

4. I suspect that the use of more examples and counter-examples in the diary will assist the respondent in reporting and classifying expenditures.

5. I would have preferred that the sponsoring

agency add some of their reactions to this study. While the authors of the paper do well exposing both the strong and weak points of the study, they nevertheless have a vested interest in the outcome. Does the sponsor think the study was successful and if so, what is next? More small scale studies? A large field test? How will an assessment be undertaken and when?

MARKER, WAKSBERG, AND ATHEY

Marker and his colleagues explore a respondent's ability to describe geographic locations in a RDD survey. Is it feasible to ask a series of questions describing the perimeter of the desired area and have the respondent classify the area correctly? The authors wish to develop efficient mechanisms for identifying geographic areas for sampling as well as areas targeted for oversampling. Their evaluation is based on comparisons of the actual physical locations with the results of the classification through a series of screener questions.

The authors describe a successful attempt at obtaining sample relevant geography in a RDD setting. Indeed, 85% of the cases that were oversampled were done so correctly, and 89% of them that were not oversampled were done so correctly. So, what's missing? The authors do not provide any insight into the meaning of the error rates. I think these error rates are good, but how much of a problem was this lack of efficiency? How costly were the errors? Were the error rates anticipated?

The nonresponse rate was about 14%. Do we have any information about the nonresponse cases? Do the authors have any evidence to believe that the nonrespondent classification rate would be the same as the respondent rate? Could the rate be worse? Are the nonrespondents more difficult to classify? I doubt it, but it would be nice to have some evidence.

Some discussion of how the questions were developed and tested would be helpful. If we believe this study is successful, other survey methods researchers would benefit from knowing the approach taken by the authors to ensure successful implementation. Finally, to what extent is the success of the work attributable to the fact that the authors were dealing with an essentially urban/rural dichotomy? The answer is probably to a large extent; however, the work does show that the approach ought to be tested in other settings.

MOELLER AND MATHIOWETZ

The paper by Moeller and Mathiowetz discusses the difficult problem of screening to provide for an oversample of the poor in the National Medical

Expenditure Survey (NMES). This is a difficult issue because the technique of screening on reported income has more than several problems associated with it. The NMES experience saw an unusually large percent (62%) of sampled persons who reported they were poor at the time of the 1986 screener, report that they were no longer poor in 1987. This rate far exceeds that found in other surveys. The authors conclude that many of the screener respondents underreported their income and, therefore, were classified incorrectly as poor at the time of the screener. The models they develop help classify individuals as to whether income was underreported at the time the screener was implemented. Furthermore, the authors go on to assert that oversampling on the basis of the characteristics in a model they develop could be more fruitful for oversampling the poor than the method of using a screener question.

Some comments:

1. I admire the tenacity of the authors in seeking a different approach to the problem--trying to identify the poor through an empirically based model. I am unconvinced that this is necessarily a better approach as it seems there is still a great deal of inefficiency in oversampling this subgroup--many cases classified as poor one year will not necessarily be poor at the time of the survey.

I would not yet give up on the "screener" question but add more questions to attempt to correctly classify the poor/non-poor status of the reporting unit.

While this paper was not specifically about the development of the screener question, I would have liked more information on how the question was formulated and tested prior to its use. The screener question is a model of brevity. Perhaps it could be improved significantly through the addition of 3-4 additional questions. Obviously, I think further testing of questions is called for before striking out on a path to predict sample members through a model.

2. Another point I'd like to raise in the context of the accurate screening of the sample is the precision at which we classify individuals. The "poor" and "near poor" criteria seem unnecessarily stringent. If the policy concern is for subgroups at depressed economic levels, then perhaps the economic criteria should be loosened.

Economic well-being does not begin once a reporting unit crosses the poverty line. I am suggesting that subgroups of policy interest can tolerate a certain amount of response error in the screener and still be of policy interest. Screening on

poverty status and ending up with a number of respondents not below poverty but at say 140% of poverty is indicative of an inefficient procedure but the results are not totally worthless. This sample could clearly enhance a number of analyses near the poverty line.

COHEN

The Cohen paper begins, at least to my knowledge, a necessary and important process for the National Center for Education Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS) program. This process should be nothing less than a systematic and comprehensive evaluation of the data collected in the IPEDS program; this includes a compilation and discussion of the sources and magnitude of sampling and nonsampling error in IPEDS. Cohen provides a useful background review of IPEDS and its scope. The error analysis he focusses on is sampling error with a few indications of nonresponse error.

The paper needs to indicate the reasons for maintaining continuity with the Higher Education General Information Survey (HEGIS). These should be explicitly stated and developed. It's not obvious to me what the political or substantive reasons are for retaining so few HEGIS institutions in the sample with certainty.

I would have liked Michael to spend more time in the paper discussing the variety of issues encountered in establishing the definition of "postsecondary education." Many of these issues affected the development of the sampling frame.

Other questions that deserve to be addressed at some time include: 1) what reliability criteria were established in order to determine sample size?; 2) can you provide some measures of the dynamic nature of the frame?; 3) what types of institutions are most subject to this dynamism?; 4) how good is the coverage and what benchmarks are used to make the assessment?

Finally, even though the paper is not about nonresponse, it is obvious that future work ought to systematically describe the nature of the nonresponse problem. Knowing something about the missing data problem will inform the NCES efforts to improve the response rate which is unsatisfactorily low.

All authors in this session should be congratulated for their efforts in providing information about error sources in surveys, and suggesting ways to reduce these error sources.