DISCUSSION

Judith M. Tanur, State University of New York
Stony Brook, New York 11794-4356

These papers by authors from BLS and Census represent reports of an ambitious program of research already spanning almost half a decade and projected to continue for another several years. It will culminate with an operational redesigned CPS, with improved questions and a careful phase-in so that the effects of changes on statistics in the labor force series will be known. I had the opportunity to "hang around" BLS during academic year 1988-89 as an NSF/ASA/BLS Senior Research Fellow, so that I got to watch some of the research in planning and in process, to listen in on some CATI/RDD interviews, and to witness some of the extensive coordination necessary to bring off a project of this magnitude and complexity across agencies and contractors. I even got to see some of the laboratory experimentation that preceded the CATI/RDD test described here, experimentation that used cognitive principles and methods. I am truly impressed both by the process and by the product.

The combination of papers presented here is fascinating. The Esposito, Campanelli, Rothgeb, and Polivka paper discusses methodology for evaluating various survey methods, and lest we fear falling into an infinite regress of considering methods for methods for methods, the Rothgeb, Polivka, Creighton, and Cohany paper gives us concrete results. I shall refer to them respectively as the methodology paper and the results paper.

What I particularly like about this pair of papers is the multiple methods brought to bear on determining which questions or series of questions should survive to the final test version of the revised CPS. Besides the earlier work in the cognitive laboratories, these investigators used behavioral coding, interviewer and respondent debriefing, and nonresponse and response analysis. They make the point that the response analysis indicates where there are differences between forms, and that the more qualitative data shed light on why such differences exist. In these evaluations, as the authors note, there is a constant tension between the aims of deciding which questions give particular trouble within a questionnaire and which of similar questions are more efficacious across questionnaires. Techniques that raise warning flags within a questionnaire (many levels of exchange, many inadequate answers, etc.) are less useful in identifying the best of several similar questions across questionnaires. These differences in function make sense if the difficulties signalled by the interactional warning flags actually refer to difficulties respondents are having with the concept itself, difficulties that persist regardless of the form of the question. Hence response analysis -- based on the premise that more reporting of a behavior is better reporting -- coupled with some wonderfully careful detective work to help to indicate when too much of a good thing is suspect -- seems to have been the main method for choosing questions for version D. As the methods paper points out, this emphasis was largely the result of the ease of quantification of the response analysis.

While my main theme here is to applaud the flexibility of this enterprise in considering all sorts of data, let me suggest that we all carry out prejudices with us. In that sense, analyst bias -- a kind of measurement error not usually considered in discussions of evaluations of survey methods -- may have played a role in the decisions made about version D. I teach in a department of sociology, and there one who uses quantitative methods is called, sometimes derisively but often with envy at the capacity to use such methods, a "quantoid." Let me suggest that we could coin a new aphorism -- "once a quantoid, always a quantoid." And in view of that new truism, perhaps we might wonder

whether some of the analytic decisions that emphasized data from the response analysis might have been different if the research team had included more people prejudiced in favor of qualitative data and practiced in using them. Perhaps such members can be included on the team for the next review.

A lovely outcome of all this research is that version D -- and presumably the revised CPS -- is more interactional, more "respondent friendly," more commonsensical, and takes far less of an Alice in Wonderland stance than words mean what the survey designers mean them to mean. For example, when it became clear that respondents' understanding of the term "on layoff" was unlikely to include the idea of expectation of recall to a job, the new version of CPS incorporated specific inquiry into such expectations. Further, for those who report themselves as retired or disabled during one month's interview, a dependent interviewing strategy that inquires only for changes in that status in subsequent months is a clear step in reducing respondent burden, as is the tailoring of subsequent questions to accommodate the retired and the disabled. Still further, asking a respondent the easiest way to report his or her income (weekly, monthly, etc.) and then asking that it be reported in that easiest way will surely reduce respondent frustration as it increases data validity. And the use of dependent interviewing for the industry and occupation questions should go a long way to reduce the spuriously high gross flow rate between industries and occupations occasioned by slight variations in reporting that result in large variations in coding, again while reducing respondent burden.

In a very carefully analyzed data set, I find I have one analytic detail that I wish had been further explored. In analyzing the vignette data from the respondent debriefing study, the results paper shows that respondents to version B were less likely than those responding to other versions to correctly classify someone working without pay in a family business as indeed working but more likely to correctly classify a volunteer workers as not working. The data are:

	A	B	C
Family worker working	47%	8%	26%
Volunteer not working	55%	92%	73%

But if we look at these same data in a slightly different way, asking the percent who said the character in the vignette was working in each case, we find the following data:

	A	B	C
Yes family worker working	47%	8%	26%
Yes, volunteer working	45%	8%	27%

The striking similarity of these percents makes one wonder if a response set is operating here and to long for a cross tabulation of responses to one vignette by responses to the other to determine if it is really the same people who are saying "yes, working" in each case.

For some years now, Steve Fienberg and I have been doing research on the parallels between experiments and surveys. Several of our papers have focussed on experiments embedded in surveys (Fienberg and Tanur 1988; Fienberg and Tanur, 1989), such experiments as the CATI/RDD experiment with its three versions of the questionnaire given to subsets of respondents. In these papers, we are not told how respondents were assigned to versions, but my presumption is that some random or pseudo random mechanism was used -- the authors are all too professional statisticians to have done otherwise. One thing Steve and I have stressed is that the design of embedded experiments should take advantage of the survey design with which it is interwoven. In particular, we have suggested that the fact that interviewers do multiple interviews means that blocking on interviewers, having each use every version of the questionnaire or other experimental treatment, would clearly increase local control and presumably decrease error variance. We have often been met with counter arguments -- either such a procedure is too complicated for interviewers who will make mistakes and add to error, or it is too likely to tempt interviewers to use a preferred procedure when a less preferred one is assigned, thus upsetting the experimental design. Although it is not explicitly stated in either of the papers that interviewers used

all versions of the CPS questionnaires, the fact that interviewer debriefing questionnaires and focus groups asked interviewers to compare versions suggests strongly to me that such blocking did indeed take place. I am delighted that these researchers saw the usefulness of having interviewers be able to compare questionnaire versions as outweighing the potential problems engendered by letting the field staff in on the experimental design. And I take their lack of mention of the occurrence of any problems of contamination through confusion or convenience to mean that there is no evidence of such contamination.

But I am still not completely satisfied. Steve and I have also strongly advocated the use of such design features as blocking on interviewers in the analysis of an embedded experiment. I can see no evidence of such use in the CATI/RDD experiment. Indeed, the methods paper points out that the interviewer data was less useful than hoped because of low power and suggests that rating scales would have provided more data and data more amenable to statistical analysis. Let me suggest that an analysis that takes into account the blocking on interviewers would also increase power on all analyses -- and further suggest that such an analysis be implemented in Phase II.

In closing, I would like to add my pleas to those of the authors for more such multi-method approaches, so we can learn more about the strengths and weaknesses of methods of evaluating survey questions as we learn more about the questions themselves. But in doing so, let us take maximum advantage of our experimental designs and let us be forceful in seeking participation from qualitative analysts.

References

Fienberg, S.E. and Tanur, J.M. (1988) From the inside out and the outside in: Combining experimental and sampling structures. Canadian Journal of Statistics, 19, pp. 135-151.

Fienberg, S.E. and Tanur, J.M. (1989) Combining Cognitive and statistical approaches to survey design. Science, 243, pp. 1017-1022.