Discussion of Papers on Cognitive Aspects of Web Survey Design

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

I feel honored to be invited to discuss these fascinating papers. And challenged!

It has been almost three decades since the start of the CASM movement. In that time, I believe some real progress has been made in understanding cognitive aspects of what we now think of as traditional surveys. We no longer merely put together catalogs of kinds of measurement errors and depend on the counsel of experience to help avoid them. We have some reasonably well-articulated theory to explain how some kinds of measurement error occur and often to give us systematic guidance of how to cope with them. The authors of the papers in today's session have played major roles in developing and systematizing that theory.

But here we go again! Before we fully understood all the measurement error issues raised by in-person interviewing, we had to cope, in addition, with new problems raised by telephone interviewing. And before we fully understood all these measurement problems, we were faced with another new set of problems with the advent of computer assisted interviewing. And now – Web surveys are presenting us with yet another set of issues. Luckily we do not have to start with catalogs of measurement errors this time. This is in part because we haven't had enough experience to amass such catalogs, and in part because we already have some theory to guide us.

That theory arises in part from CASM, but also from considerations of the survey interview as social interaction and from such fields as the study of visual perception and visual cognition. And we know a lot of the variables that we should be considering – properties of images and their relevance, issues of spacing, aids to comprehension, and so forth.

Thus, for example, when Mick Couper and his colleagues explore the effect of a picture of a sick woman or a fit one on self-ratings of health, they can evoke the theory dealing with assimilation and contrast effects. And they can evoke it not only as a post-hoc explanation of experimental outcomes, but perhaps even to predict what new experiments should show. Similarly, Lucy Suchman and Gitti Jordan showed us that an interviewer who, in the interests of standardization, refuses to offer clarification not only causes interactional troubles in an interview, but collects misleading data as well. In response, and under the CASM umbrella, Fred Conrad (together especially with Michael Schober and other colleagues) has been exploring the advantages and costs of conversational interviewing. These concerns readily transfer to web surveys – and Fred and his colleagues have presented a paper exploring the effects of different sorts of feedback and clarification to respondents (Rs) taking web surveys. And Jon Krosnik can bring his ideas of survey satisficing – Rs expending only a minimal effort – to the examination of web surveys.

Let me turn briefly to the papers themselves. Jon Krosnik has given us a straightforward and careful comparison of data quality between an RDD survey carried out by a university survey center, and two kinds of web surveys – one based on an RDD sample, though Rs have to opt in at several stages, and the other a fully volunteer recruitment. The study also tackled what I see as the basic problem of web surveys, our inability to draw a true probability sample of the population to which we would like to generalize. Not only are there coverage problems – not everyone has web access – but there are also issues of self-selection absent a really good frame. Very surprisingly, Jon – with LinChait Chang – finds the panel recruited by Knowledge Networks no further from the demographic profile of the March CPS than was the RDD sample. The volunteer survey from Harris Interactive was less accurate. The trouble with these findings, as with any nonprobability method, is that we never know when a procedure that has worked reasonably well in the past will, under special unforeseen circumstances, suddenly fail. Perhaps we should be cheered by the fact that all three procedures underestimated voter turnout, as we have come to expect.

In the work reported here, on several dimensions Jon finds differences in data quality across modes. He examines random measurement error, survey satisficing, and social desirability response bias. Harris Interactive (volunteer panel) had higher data quality than Knowledge Networks (web survey based on a RDD sample), which had higher quality than the RDD sample carried out by the university survey center. I find especially interesting the reasoning behind the measurement of random error. Jon looked for a systematic relationship between predictors of vote intention (41 of them, measured at 2 points in time) and actual post-election report of whom the R

voted for. The reasoning is that the greater the systematic variation measured by logistic regression, the less random error there is in the system. This is clever and sophisticated – but I would dearly love to see a record check survey conducted to help nail down the issue of data quality.

To see if these differences are really attributable to mode – rather than the myriad other things that might vary across these samples – Jon reports a follow-up laboratory experiment, where subjects were randomly assigned to simulated phone interviews via an intercom or to take the survey on a computer. The questions were the same as in the post election field survey. Concurrent validity was higher for the computer than for the intercom, especially for low-SAT Rs. Satisficing took place more frequently on the intercom than on the computer, especially too for low SAT Rs. A recency effect occurred on the intercom but not on the computer– and this effect was stronger for high SAT Rs. So the effect of cognitive ability is complicated. There was again less social desirability response bias for the computer survey than for the intercom. Surprisingly, the intercom interviews took longer than the computer interviews, so the improvement in response quality was not attributable to more time on task.

I especially liked the proposal for future research to track down *why* the computer seems better – Are the interviewers too unstandardized? Is it because Rs work at their own pace? (Pacing issues on a web survey are considered in the paper by Fred Conrad.) Is because there are reduced working memory demands because of visual presentation? But Jon raises the questions in order to try to improve *phone* surveys.

The paper Fred Conrad presented (reflecting work with Mick Couper and Roger Tourangeau) looks at the special features of web surveys that make (or can make) them interactive as opposed to the static – one size has to be engineered to fit all and the R is responsible for branching correctly – confines of PAPIl surveys, perhaps the web survey's nearest analog.

The first of these interactive aspects they consider is the usefulness of giving Rs on-going feedback about his/her progress through the survey. (In a paper and pencil questionnaire Rs can – although we often don't want them to – skip ahead and see how much more is coming.) The basic question is whether feedback can be designed to encourage (You're making good progress – the end is fast approaching!) or will it be discouraging (OMIGAWD, there's so much more to do!).

They varied the report of progress by making it reflect (1) the proportion of screens completed, yielding a straight line progress indicator, (2) the log of the current screen as a proportion of the log of the final screen, yielding a progress indicator that went from faster to slower, and (3) the inverse log of the current screen as a proportion of the inverse log of the final screen, yielding a progress indicator that went from slower to faster. They found that early encouragement (faster to slower) helps – there were fewer and later break-offs, the survey was judged more interesting, and was judged to take less time. Rs in the slower-to-faster condition, however, on average judged the survey more useful (perhaps an example of cognitive dissonance?) although less accurate. The authors point out that these Rs were volunteers – the effect might be even stronger with less motivated Rs. The study was not done in conjunction with a prior time estimate – how would such an estimate change things? Would it make the progress reporting unnecessary?

The authors raise an ethical question – is it right to deny progress information to Rs if, as on a web survey, it's available. Is it ethical to provide a variable speed indicator, even if it motivate Rs and leaves them feeling better about the experience? I would put the matter more strongly. Since nobody calculates log completed/log total as percent finished, this may be a potentially dangerous deception. What long run effect might it have, when Rs, especially on web panels, take repeated surveys and find the feedback they get is not congruent with their experiential gut feelings? Will they come to doubt or discount the progress reports? Will such discontent spill over onto other aspects of the survey? I think we've gained valuable information; I also think we need to be careful how we use it.

The other research Fred reports on is in parallel with his and Mike Schober's work on conversational interviewing. How can we provide definitions to Rs so that their comprehension of our questions will be congruent with what we intended? They varied the difficulty of getting a definition, how badly it was needed, and its usefulness. They found that definitions were rarely used, were used more often for technical terms than for commonly used words and "Difficulty of obtaining a definition reduced its later use." They report that of those accessing one or more definitions, 56% were in the 1-click treatment, 24% in the 2-click treatment, and 20% in the click and scroll treatment. I would like to see what percent in each condition asked **again** for a definition, given that they had asked at least once.

They also worked on making the web instrument more sensitive to Rs' need for definitions. They contrasted: no clarification, user initiated clarification, clarification automatically coming after a fixed interval without a click, and clarification automatically coming after a fixed interval without a click but using a longer interval for those over 65 than for those under 35, and clarification always present. They used scenarios so that they could determine if correct answers were given. Easy questions were always answered right; complicated questions were more likely to be answered right when clarified. The group based clarification worked well for the younger people, giving as much accuracy as the condition when clarification was always present, but not for older people, where the group based accuracy was considerably less than when clarification was always present and was even less than the generic clarification condition. The older Rs actually got more clarification in the generic condition (when presumably it came faster) than in the group-based clarification condition. The authors point out that sometimes the clarification came after the older people had already answered. So, as an over-65, I suggest that perhaps we're not as slow as you youngsters think. Or perhaps we're more variable. The idea of group based feedback seems a good one – and something that could only be done on web surveys. But perhaps we need to define groups better. The authors suggest other groupings – perhaps based on computer experience. I wondered if everyone knew how to click on the hyperlinks that brought up the definitions.

Mick Couper, with the same co-authors, tackles problems of imagery on web surveys. They point out the images can have positive and negative effects. They consider the visual context and point out that if an image is close to a question the viewer assumes they must be related and works hard to understand how and to try to make sense. Images may direct attention, add explanation, enhance recall, activate mood, and trigger attitudes through priming – that is, they may affect all parts of the survey responding process. And images may be perceived differently by different Rs. All this is terribly complicated. [At this point a DVD "Surprising Studies of Visual Awaareness" prepared by Daniel Simons and produced by VisCog Productions what shown. It demonstrated the effects of selective attention.]

Mick recounts experiments on using the visual organization of the screen to impart information to the R. One was on separating substantive and non-substantive answer choices by a horizontal line or a space. Fewer non-substantive answers were given when there was visual separation – Rs were making sense of the organization of the screen. Similarly, response options unevenly spaced horizontally affected Rs' choices. When the endpoints of the scale were "way out" the mean deviation increased; and when the spacing between responses increased with increasing response, Rs tended to move their responses closer to the visual middle of the scale rather than the conceptual middle.

The authors did an experiment using high and low frequency exemplars of behavior they were seeking frequency reports of. They hypothesized that the presentation of high frequency exemplars would yield more reporting. This turned out to be true, but the results were very complicated – there was priming too, with Rs using the images as cues about what to include. The images seemed to function to give much more information than the experimenters had counted on. This is an example of how much we have to learn, and the difficulty of calibrating the effects of images. The authors point out that the images had no effect on break-offs, so their advice is to use them with caution, not just for visual appeal. I couldn't agree more.

Another experiment was on contrast effects. How does the image of a sick woman versus that of a fit and jogging woman affect self-ratings of health? The authors found an expected contrast effect: higher health ratings in the presence of the image of the sick woman than in the presence of the image of the fit woman. But when the image was in the header of the question there was an assimilation effect, with health rated higher in the presence of the "fit" image than in the presence of the "sick" image. Wow! And besides, the contrast effect works only for one scale (ranging from excellent to poor) but not for the other (extremely good to extremely poor). Double Wow!

And then there's the experiment on priming – does looking at a picture of a storm affect one's mood? Yes, but only on one question about mood (how often do you feel blue?) and only when that question is in close proximity to the picture. Wow again!

I applaud this program to understand visual images, visual perception, and visual cognition. There is no doubt that Rs notice images (at least *sometimes*) and that images can *sometimes* affect answers. Rs use spacing and relative position as cues to the meaning of answers. So we have to carry out this research. But I worry enormously about all the variables in all their combinations that these authors have to consider – and I envy them enormously for the broad areas they have available to explore.