DISCUSSION: PAPERS ON INTERDISCIPLINARY RESEARCH INVOLVING THE COMPUTER AND COGNITIVE SCIENCES

Gordon B. Willis, Research Triangle Institute
Suite 420, 6110 Executive Blvd., Rockville, MD 20852

Key Words: Cognitive Sciences, Survey Methods

The preceding papers represent the newest generation of developments in the area of survey methodology that center on either increasing data quality or reacting to ongoing changes in the world of survey-taking. Below I provide a critique of each paper and suggest areas for continued research efforts.

Graesser, Wiemer-Hastings, Wiemer-Hastings, and Kreuz

The authors describe a computer-based tool (QUAID) for detecting problems in survey questions. QUAID is intended as an intensive screening device that will save the questionnaire designer effort by identifying suspect questions. To their credit, the authors make efforts to specify what they mean by the presence of a "problem" in a survey question, and focus on six defined problem varieties that QUAID is argued to be capable of uncovering.

One reaction I have is that the paper could focus more on the increasing number of survey problem classification schemes that already exist in the literature. Their model does appear to be consistent with these. It may be that at this point, the "four corners of the globe" have been discovered, with respect to classification of problems existing in commonly administered survey questions, and that this area is best viewed as one requiring research synthesis and metanalysis, as opposed to further generation of additional models.

The overall ambition of the Graesser et al. paper is to empirically test QUAID against other forms of questionnaire pretesting that purport to detect survey questions containing problems. The authors focus especially on evaluation by experts through either what might be termed "armchair review," or else that based on cognitive interviewing or behavior coding. They suggest that review by experts is unreliable, and that a tool such as QUAID is therefore needed. However, note that although expert armchair-based review may well be idiosyncratic, this fact does not imply, by extension, that experts' use of either cognitive interviewing techniques or behavior coding is also unreliable. Significantly, the authors do not cite the accumulating body of evidence devoted to assessing the effectiveness of cognitive interviewing techniques and other pretesting methods. Hence, Graesser et al. provide no general comparison of QUAID to techniques other than the most simple form of review that involves human experts.

Still, the comparison of the QUAID system with human "expert review" is enlightening. Using signal-detection analysis, the authors find that, in comparison with experts, QUAID produces high sensitivity, yet low specificity (comparatively many problems are identified by QUAID). This finding is notable, in that previous assessments of expert review have sometimes found that experts tend to apply problem codes more frequently than do other pretesting methods. The clear question would then seem to be whether expert review can itself be considered a "gold standard" by which to assess QUAID or any other approach to questionnaire pretesting. I suggest that this is an unresolved issue, but that the method that is "best" in a particular circumstance may be the one which effectively balances sensitivity (finding problems that exist) with specificity (appropriately failing to flag problems when they *don't* exist). Graesser et al. do appear to agree that it is not clear that a particular gold standard now exists by which to anoint any evaluated procedure as either adequate or lacking.

In final analysis, it may be that there is no one best pretesting procedure; those that are routinely used may consist of a spectrum of techniques that do not compete, but rather complement one another as they are applied through different phases or steps in the questionnaire development process. Focus groups are useful early in the concept development phase; expert review (or QUAID) can efficiently be applied to a first draft, cognitive interviewing to later drafts, and behavior coding to a version that is "field-test ready." There is no single technique that should supplant the others.

Finally, note that there may be inherent limitations to computer-based question evaluation systems; in particular, QUAID is appropriate for *question* evaluation, as opposed to the more general review of the total *questionnaire*, which is often of fundamental interest. Further, QUAID currently does not account for the fact that survey questions are often targeted toward particular population groups (as an example, a medical term that is too technical for a survey of the elderly may work well for a survey of physicians). However, despite these limitations, it seems that QUAID is likely to be a useful component of the pretesting toolbox; it is sensible to attempt to rely on the increasing sophistication of computer-based systems, especially as these are found to be capable of representing natural-language processes, and to expect these to shoulder a good deal of the burden of preliminary assessment of survey questions.

Rips, Conrad, and Fricker

Rips and his colleagues make a significant contribution simply by providing a careful and coherent description of the Seam Effect in panel surveys. The authors further propose a cognitive mechanism underlying this effect— a combination of memory and guessing/estimation (the latter associated with the desire to be consistent within the interview, or the "constant wave effect"). Rips et al. then endeavor to test this hypothesized mechanism through the use of a simulated panel survey in which critical elements of the design can be controlled. The key to their procedure is the use of "implanted memories"; by testing memory for a known attribute (the actual content of a previously presented questionnaire), they take control of the encoding of events, rather than leaving this to the vagaries of real life. Through use of this imaginative design, the authors are able to produce a synthetic Seam Effect under conditions that they propose will do so, and a much reduced effect under conditions designed to ameliorate it. It can then be argued that the effect has been adequately explained in cognitive terms.

Of course, a critical potential danger in the use of a synthetic or arguably contrived experiment is that the effect studied "under the microscope" differs fundamentally from the applied phenomenon that is the ultimate focus of the study. In this case, it is possible that memory for questionnaire content may operate differently from that underlying the answering of actual survey questions. However, this would not seem to be an overwhelming criticism of the Rips et al. findings; it is difficult to believe that there are two cognitive mechanisms that produce Seam Effects, one applying to real surveys, and the other to artificial ones. I would venture to guess that the investigators have in this case focused their search in the right place, as opposed to simply where the light is brightest.

Beyond explaining a phenomenon that may lead to response error, Rips and his co-authors further endeavor to moderate this effect by modifying controllable survey conditions. Interestingly, they find that the Seam Effect is minimized through the use of recall-by-time, rather than by-topic, and through use of backward rather than forward recall. They further point out that the optimal procedure (recall by month, backward) is not typically applied in large-scale Federal surveys, which suggests that we may be relying on inferior, yet improvable procedures.

It might be argued that the perspective offered by this particular research effort is necessarily myopic, in that it recognizes only the existence of the Seam Effect as a source of questionnaire-based error in panel surveys. Many factors contribute ultimately to questionnaire-design decisions, and there may be countering factors which select against the use of the advocated procedure. However, I conclude that Rips and his colleagues have made a valuable contribution by better illuminating one source of response error which can be addressed through design decisions.

Belli

Similarly to Rips, Conrad, and Fricker, Belli focuses on survey designs that require the autobiographical recall of events. He suggests that the Event History Calendar (EHC) approach has been found to be useful for these tasks, in comparison with a linear list of questions (a "Q-List") obtaining identical information. Further, he proposes that a novel, computerized EHC will be effective, because the computer can handle tasks more efficiently than paper-and-pencil, can "keep track of" the interviewer, and in particular, provides a means for maximizing flexibility with respect to the interviewing approach.

Several issues come to mind concerning the application of Belli's system. First, he states that respondents reported the same level of subjective burden for EHC as for the Q-list. This finding begs a related question concerning the *objective* degree of burden, compared to a standard (Q-List) method; it is sometimes found that increasingly intensive question-administration procedures are more accurate, but at the cost of markedly increasing survey administration time. Demonstrating that the flexible computerized EHC system does not significantly increase time burden would be a positive finding, especially in a climate in which respondents (as well as OMB) are motivated to limit the length (duration) of survey questionnaires. A second consideration relevant to computerized event history calendars is training costs; once interviewers are given flexibility, how difficult or time-consuming is it to instruct them to perform the various tasks efficiently?

However, the most significant feature of the computerized EHC may be its focus on flexible interviewing. Debate currently rages in the survey methodology literature concerning whether we should be moving towards greater flexibility, as opposed to greater standardization, of interviewer behavior. Thus, the full utility of Belli's system may be realized to the extent that the flexibility it provides is found to enhance critical features of the interview associated with either burden or data quality.

Tourangeau, Couper, Tortora, and Steiger

Researchers at the University of Maryland and at Gallup, Inc. have collaborated to investigate pressing issues that emerge as survey researchers increasingly look to the Internet as a mechanism for the administration of population surveys. They ask a compelling question: Do the increasing opportunities to incorporate elements that render the computer interface as human-like in sound and appearance represent an improvement, or rather a source of potential error?

Tourangeau and colleagues operationalize "humanness" of the interface, or what they label Social Presence. They conclude that although their investigation did not find the effects of Social Presence to be potent, the anthropomorphism of the machine, in the guise of features that make the survey more attractive to the potential respondent, may influence responding in critical ways.

My first question would be whether the authors have found evidence that the features they embedded did in fact make the survey more attractive. For example, did the more human-like interface lead to a reduction in the number of break-offs during the interview? One underlying issue may be the degree of attractiveness (or "sales value") of the survey that is achieved through enhanced Social Presence; a second might be the subsequent effects on response tendencies, as the individual completes the Web-based interview. These effects may even function in opposition, with human qualities that spur potential respondents to participate also having the undesired effect of adversely influencing responses to survey questions.

A second reaction relates to the obtained response rate, reported as 20% for the larger of their two studies. This is an extremely low value, relative to usual survey response rates, and it is possible that a sample bias existed in which those taking the Web survey were already highly motivated, and therefore not likely to be much affected by the manipulated variables. Repetition of the experiment, using a more general respondent audience, could be illuminating.

Finally, it seems that the finding that "personalization" influenced responding may not represent a manifestation of the effects of the novel phenomenon of Social Presence associated with use of a computerized interface, but rather a traditional psychological demand effect; whether administration is by computer, paper-and-pencil, or some other means, it has often been found that once respondents are aware of features of the investigator which may be related to research hypotheses (such as gender), they modify their performance accordingly. Therefore, a specific focus on demand effects as an explanatory factor would be a reasonable next step.

Most generally, the possibility exists that Social Presence subsumes several variables, one directly related to the extent to which the interface is present in human-like form, and another related to the degree to which the respondent is made aware of relevant characteristics of the investigator (as opposed to the computer "interviewer"). A number of subtly different factors could presumably have varying effects on survey respondents, and combine to create an overall Social Presence Effect.

Summary

Looking across papers serves to identify two commonalities or general themes:

(1) ***Standardization:*** Belli advocates increased interviewer flexibility, whereas Tourangeau et al. cite increased standardization as a positive feature of self-administered Web-based surveys. The papers represent two sides of the coin associated with a more general debate surrounding the use of standardized interviewing. The issue is unresolved, but as for other issues impinging on survey design, one possibility is that both papers are correct; under some circumstances, flexibility is advantageous; under others, one should strive for maximal standardization. The next meaningful step would be for methodologists to better specify these conditions.

(2) ***Adaptation to respondent behavior:*** All four papers represent attempts to understand how respondents behave, and then to do something about it, whether from the perspective of question design (Graesser at al.), more general questionnaire design/organization (Rips et al.; Belli), or administration features that serve as context and background to the questionnaire (Tourangeau et al.). The authors in this session appear to be implicitly resigned to the fact that we can't change respondents--we can only change our own approach. That is, contrary to a tradition suggesting that we train respondents, on-the-job, to behave as we desire, we instead acknowledge that we have limited control over the myriad ways in which those respondents react to our presented materials. So, we had better understand the dominant trends that influence response tendencies, and adjust for them in a way that maximizes the utility of whatever it is that we are getting back.