COGNITIVE ISSUES IN THE DESIGN OF WEB SURVEYS

Roger Tourangeau & Mick P. Couper,
Survey Research Center, University of Michigan
Robert Tortora and Darby Miller Steiger, The Gallup Organization
Roger Tourangeau, Joint Program in Survey Methodology, 1218 LeFrak Hall,
University of Maryland, College Park Maryland 20742

KEY WORDS: Computer Administration, Web Surveys, Mode Effects, Deference

1. Introduction

The development of new methods for collecting survey data—particularly Web surveys and administration of recorded questions by telephone—may be ushering in a golden age for self-administered surveys. The new methods of data collection seem to combine the power and complexity of computerization with the privacy of self administration. At the same time, because they do not require an interviewer, they may reduce other types of survey errors and could dramatically lower the costs of collecting survey data.

The evidence from the survey literature overwhelmingly suggests that self administration improves the reporting of sensitive behaviors, such as illicit drug use. For example, six similar studies have compared reports of illicit drug use under interviewer and self administration of the questions (see Tourangeau and Smith, 1998). Across different drugs, different time frames, and the different studies, self-administered questions yielded a median increase of 30 percent in the proportion of respondents reporting they had used illicit drugs. In some cases, the gains from self administration were quite dramatic. Other studies have demonstrated the advantages of self administration for reports about sexual behavior (Boekeloo, Schiavo, Rabin, Conlon, Jordan, and Mundt, 1994; Tourangeau and Smith, 1996; Tourangeau et al., 1997; Turner et al., 1998), alcohol consumption (Aquilino and LoSciuto, 1990; Hochstim, 1967), abortions (Lessler and O'Reilly, 1997; London and Williams, 1990; Mosher and Duffer, 1994; Mott, 1985), and church attendance (Presser and Stinson, 1998).

Aside from its impact on the reporting of potentially embarrassing information, self administration offers other advantages. Interviewers make errors in reading the questions, their inflections or tones suggest certain answers, and they affect the respondents' answers in other subtle (or not-so-subtle) ways. Moreover, when the questions are related to obvious characteristics of the interviewers (such as their race or sex), respondents sometimes may alter their answers to avoid offending the interviewers (Anderson, Silver, and Abramson, 1988; Hatchett and Schuman, 1975-1976; Kane and Macaulay, 1993; Schuman and Converse, 1971). Such variations across interviewers threaten the validity of survey findings. Self administration ensures greater standardization of delivery of the questions, reducing or eliminating interviewer effects (e.g., Tourangeau et al., 1997).

1.1 Variations across Methods of Self Administration

The gains from the new computer-assisted survey methods—reduced social desirability bias, greater standardization, reduced interviewer variation, and decreased cost—may not be intrinsic to these methods but may depend on relatively peripheral features of the technologies or of the interface used in specific applications. Most survey researchers believe that self administration yields more accurate answers to sensitive questions because it is more private than interviewer administration; it eliminates the need for the respondents to make embarrassing admissions directly to another person. There is, however, mounting evidence that different methods of self administration can produce systematic differences in the results. For example, a recent study by Beebe and his colleagues (Beebe, Harrison, McRae, Anderson, and Fulkerson, 1998) showed that, in a school setting, respondents were more likely to admit to illicit drug use, fighting, and other sensitive behaviors on a paper self-administered questionnaire (SAQ) than on a computer-administered version of the same questions. Beebe and his colleagues argue that various features of the setting for the computer administration can reduce the sense of privacy and, as a result, affect the answers. In their study, the computerized questions were administered in the school's computer lab on networked computers via terminals that were physically close to one another. Any of these features may have reduced the apparent privacy of the data collection process. These and similar results suggest that an electronic questionnaire is not inherently more private than a paper version of the same questions.

1.2 The Role of Social Presence

Thus, although the survey evidence suggests that computerized self administration reduces reporting errors (at least as compared to interviewer administration), the impact of computerization itself is less clear. Several studies suggest that the impact of computerization

depends on characteristics of the respondents, the setting, and the interface between the program and its users.

In addition, a growing body of research suggests that relatively subtle cues (such as "gendered" text or simple inanimate line drawings of a face) in a computer interface can evoke reactions similar to those produced by an interviewer, including social desirability effects. Nass, Moon and Green (1997), for example, conclude that the tendency to stereotype by gender can be triggered by such minimal cues as the voice on a computer. Based on the results of a series of experiments that varied a number of cues in computer tutoring and other tasks, Nass and his colleagues (Fogg and Nass, 1997; Nass, Fogg, and Moon, 1996; Reeves and Nass, 1996) argue that computer interfaces (even the words used in a text-based tutoring task) can engender reactions from subjects similar to those evoked by interactions with other people. Their central thesis is that people treat computers as social actors not as inanimate tools (see also the review by Couper, 1998).

Additional support for the hypothesis that a computer interface can function as a virtual human presence comes from a study by Walker, Sproull, and Subramani (1994). They administered questionnaires to people using either a text display or one of two talking-face displays to ask the questions. Those interacting with a talking-face display spent more time, made fewer mistakes, and wrote more comments than did people interacting with the text display. However, people who interacted with the more expressive face liked the face and the experience less than those who interacted with the less expressive face. In a subsequent experiment, Sproull and his colleagues (1996) varied the expression of a talking face on a computer-administered career counseling interview; one face was stern, the other more pleasant. The faces were computer-generated images with animated mouths. They found that:

> People respond to a talking-face display differently than to a text display. They attribute some personality attributes to the faces differently than to a text display. They report themselves to be more aroused (less relaxed, less confident). They present themselves in a more positive light to the talking-face displays. (p. 116)

Thus, the addition of a variety of humanizing visual and/or aural cues, as is possible in Web surveys, may negate or at least mitigate the beneficial effects of self administration, especially for items of a sensitive nature.

In summary, some of the most intriguing findings concerning the differences between modes of self-administered data collection seem to reflect a variable we call *social presence*. To the extent that the method of data collection, the data collection setting, or the interface gives the respondent a sense of interacting with another person, it may trigger motivations similar to those triggered by an interviewer. These motivations include the desire to avoid embarrassing oneself or giving offense to someone else, as well as enhanced motivation to complete the interview. Web surveys offer ample resources for attracting the interest of the respondent, but even apparently innocuous characteristics of the interface can create a sense of social presence, producing social desirability and related response effects.

We carried out two Web experiments that examined the impact of social presence. Both experiments manipulated two features of the interface between the respondent and the electronic questionnaire. One was the degree that the program seemed to interact directly with the respondent, for example, using the respondent's name or repeating the answers the respondent had just provided. The other was the degree the interface was personalized; we personalized the interface by having the program display pictures of one of the researchers, along with personalizing messages ("Hi! My name is Roger Tourangeau. I'm one of the investigators on this project."). We thought that making the survey more interactive and personalizing it would increase the sense of social presence. Increasing the level of social presence would, in turn, reduce the respondents' willingness to provide candid answers to sensitive questions or to take positions the researcher might find offensive.

2. Methods

We carried out two studies that examined the impact of characteristics of the interface on the responses obtained in a Web survey. Our first study compared six versions of a Web survey administered to 202 participants in a Gallup Web panel. The next study compared the same six versions of the survey in much larger sample of Web users purchased from a commercial vendor, Survey Sampling, Inc. (SSI).

2.1 Study 1: Gallup's Web Panel

Questionnaires. The different versions of the Web questionnaire differed along two dimensions—the degree that the program seemed to interact with the respondent and the degree that it presented personalizing cues. The high interaction versions of the questionnaire used the first person in introductions and transitional phrases (e.g., "Now I'd like to ask you a few questions about the roles of men and women") and occasionally echoed back to the respondents their earlier answers ("According to your responses, you exercise once daily ..."). The low interaction versions used more impersonal language ("The next series of questions is about the roles of men and women") and gave less tailored feedback ("Thank

you for this information"). At several points in the questionnaire, the personalized versions of the questionnaire displayed a picture of one of the male researchers, one of the female researchers, or the logo for the study. Along with the investigator's picture, the program displayed relevant statements from the investigator: "Hi! My name is Roger Tourangeau. I'm one of the investigators on this project. Thanks for taking part in my study."

The level of interaction variable was crossed with the personalizing cues variable, yielding the six versions of the questionnaire. All six versions included the same items:

- Ten items on gender attitudes (taken from Kane and Macaulay, 1993);
- Three items on diet and exercise;
- Five items on drinking and illicit drug use;
- 16 items from the Marlowe-Crowne Social Desirability scale (Crowne and Marlowe, 1964);
- 19 items from the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984);
- Questions about voting and attendance at church;
- Three items on trust;
- Nine debriefing questions;
- Demographic questions.

We included the gender attitude items to see whether our attempt to personalize the interface produced effects paralleling the gender-of-interviewer effects with actual interviewers—that is, more pro-feminist responses with the "female" than with the "male" interface. The items on diet, exercise, drinking, drug use, voting, and attendance at church were all included to test the hypothesis that humanizing the interface (both by personalizing it and by making it more interactive) would increase the number of socially desirable responses. The Marlowe-Crowne items and the BIDR have been used for similar purposes (to measure socially desirable responding) in the work by Nass and his colleagues, and we included them in our studies for the sake of comparability. We included the trust items to see whether the impact of the experimental variables was greater among those low in trust. (A study by Aquilino and LoSciuto, 1990, suggested that those who are low in trust are more sensitive to the mode of data collection than those who are high in trust.)

On average, the questionnaire took about 15 minutes to complete.

Sample. The Gallup panel consisted of a adults who connect to the World Wide Web from home. These Web users were identified from a national stratified RDD sample. The sample design divided all telephone exchanges into two strata based on census income figures. One stratum included households with estimated annual incomes of $50,000 or less; the other stratum included households with estimated incomes higher than $50,000. This stratification took advantage of the higher incidence of Web users among households with higher income. Households in the RDD sample were contacted by telephone and screened for eligible adults—those that accessed the Web from home in the last 30 days. (In households with more than one eligible adult, one was randomly selected for the panel.) Besides screening for eligibility, the telephone survey collected data about each respondent and attempted to persuade them to become panel participants by completing one Web survey a month for six months. Incentives were offered to boost participation in the panel.

2.2 Study 2: SSI Sample

Because of the small size of the Gallup Web panel, we decided to replicate the findings with a larger sample. We used the same six versions of the questionnaire as in Study 1.

The frame for the SSI sample consists of more than seven million e-mail addresses of Web users. SSI has compiled this list from various sources; in each case, visitors to specific Web sites agreed to receive messages on a topic of interest. SSI selected a sample of 15,000 e-mail addresses and sent out an initial e-mail invitation to take part in "a study of attitudes and lifestyles." The e-mail invitation included the URL of the Web site where our Web survey resided and a PIN number (which prevented respondents from completing the questionnaire more than once). After ten days, SSI sent a second e-mail prompting cases who had not yet completed the survey. A total of 3,047 sample members completed the questionnaire, for a response rate of approximately 20 percent. (Less than one percent of the e-mails bounced back as invalid addresses.) Another 434 began the survey but broke off without finishing it. We focus here on the respondents who completed the survey.

3. Results

In analyzing the results from both samples, we grouped the items into two main categories—gender-related attitudes and sensitive items. The sensitive items included the questions on drinking and illicit drug use, on diet and exercise, and on voting and church attendance. For the gender attitude items, we created a scale that combined responses across the ten items, by scoring responses to each item in a consistent direction and then averaging across the items. Similarly, we created an index to combine answers to a number of the sensitive questions. Our index was the number of embarrassing answers given in response to those questions. This index varied from 0 to 7. Respondents got a point each if they reported they consumed more dietary fat than the average

person, were 20 pounds or more over their ideal weight, drank alcohol almost every day (or more often), had smoked marijuana, had used cocaine, did not vote in the last election, and did not attend church in the last week. In addition to the gender attitudes and sensitive admissions scales, we examined respondents' Marlowe-Crowne Social Desirability scores and their scores on the BIDR items.

Overall findings. Table 2 shows the results by experimental group and study for the gender and sensitive items scale and for two of the individual sensitive items—on voting and marijuana use. (We show results for the latter two items just to give a better feel for the results.) In both studies, neither of the variations in the interface had much impact on reporting on the sensitive items. There were no significant effects in either study on the sensitive admissions scale, on the Marlowe-Crowne SD scores, or on the BIDR scores. There were a few scattered findings for some of the individual sensitive items. For example, for reports about voting in Study 2, the personalization variable had a significant impact (χ^2=6.35, df=2, p <.05). As expected, the respondents who got the least personalized versions of the survey (which displayed the logo rather than pictures of either investigator) were least likely to say they had voted in the most recent election. In general, though, neither the level of personalization nor the level of interaction had much effect on reports about sensitive topics.

By contrast, the personalization variable did seem to affect reported gender attitudes. In both studies, we expected respondents of both sexes to report the most pro-feminist attitudes when the program displayed pictures and messages from the female investigator and the least pro-feminist attitudes when the program displayed the pictures and messages from the male investigator. We expected the group who got the survey logo to fall in between the other two. This pattern was apparent in both studies, and it was significant in Study 2 (F=5.52, df=1,3028, p<.05) and marginally significant in Study 1 (F=2.41, df=1,196, p<.12).

Interactions with other variables. Our results—especially those in Study 1—were much weaker than the ones reported by Nass, Sproull, or their colleagues. We were puzzled by the discrepancy. We used some of the same measures as the past work (e.g., the BIDR), and our sample sizes were larger (and our manipulations of the interface more blatant) than in the earlier studies. The prior work had largely been carried out with college students in laboratory settings. In addition, the respondents in our first study were panel participants with considerable prior Web survey experience. In Study 2 (where we had a large sample), we examined several variables—whether the respondent was currently a student, age, prior survey experience, and level of trust—that we though might interact with the experimental variables and explain why our results differ from those of the earlier studies. For example, we tested the hypotheses that students are more sensitive to the characteristics of the interface and that respondents with prior experience with Web surveys would be less sensitive to them. None of these hypotheses received much support—we did not find any significant interactions between these individual differences variables and the experimental variables on the reporting of sensitive information or gender attitudes.

The characteristics of the interface we varied did not have much effect on reports about sensitive topics, like voting, drinking, and illicit drug use. We thought that personalizing the interface and making it more interactive would make responding to a Web survey more like taking part in a face-to-face interview and less like filling out a paper questionnaire. Across our two studies, we found little evidence that this was so for the sensitive questions. On the other hand, we did find that respondents seemed

Table 1. Results (and Sample Sizes), by Condition and Study

	Gender Attitudes		Sensitive Admissions		% Voted in Last Election		% Smoked Marijuana in Last Year	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
Low interaction	.08 (97)	.25 (1522)	2.50 (97)	3.24 (1522)	80.4 (97)	53.2 (1522)	8.3 (97)	10.7 (1473)
High interaction	.17 (105)	.24 (1520)	2.81 (105)	3.28 (1523)	77.1 (105)	52.2 (1520)	9.9 (105)	10.2 (1487)
Logo	.10 (56)	.25 (993)	2.57 (56)	3.27 (994)	71.4 (56)	52.8 (993)	5.6 (54)	10.8 (959)
Male Picture	.08 (75)	.21 (1052)	2.55 (75)	3.21 (1058)	84.0 (75)	55.3 (1058)	9.7 (75)	9.9 (1037)
Female Picture	.19 (71)	.27 (993)	2.85 (71)	3.31 (993)	78.9 (71)	49.7 (993)	11.3 (71)	10.5 (964)

Note: Higher numbers on the gender attitudes scale indicate more pro-feminist responses. The sensitive admissions scale ranges from 0 to 7. Parenthetical entries are cell sizes.

to be sensitive to the "sex" of the interface in answering gender-related attitude questions. When the Web survey displayed messages from a female investigator along with her picture, respondents reported more pro-feminist attitudes than when the program displayed a male investigator's picture and messages. Respondents in the "ungendered" condition (who got neither set of pictures and messages) came out between the other two groups. Thus, the respondents did display *some* sensitivity to the interface in formulating their answers.

4. Discussion

The characteristics of the interface we varied did not have much effect on reports about sensitive topics, like voting, drinking, and illicit drug use. We thought that personalizing the interface and making it more interactive would make responding to a Web survey more like taking part in a face-to-face interview and less like filling out a paper questionnaire. Across our two studies, we found little evidence that this was so for the sensitive questions. On the other hand, we did find that respondents seemed to be sensitive to the "sex" of the interface in answering gender-related attitude questions. When the Web survey displayed messages from a female investigator along with her picture, respondents reported more pro-feminist attitudes than when the program displayed a male investigator's picture and messages. Respondents in the "ungendered" condition (who got neither set of pictures and messages) came out between the other two groups. Thus, the respondents did display *some* sensitivity to the interface in formulating their answers.

One other bit of evidence suggests that our variations in the interface affected reactions to the Web survey. We included debriefing items that asked respondents to rate how much completing the survey was like dealing with a machine and how much it was like interacting with a computer. Responses to these two items were highly correlated, and we combined them. In Study 1, the respondents who completed the low interaction version of the survey found it more machine-like than those who completed the high interaction version (F=2.89, df=1,195, p<.10). In Study 2, there were significant variations in ratings on the same scale depending on which picture the Web survey displayed (F=3.45, df=2,3013, p<.10).

It will be tempting to capitalize on the power of the Web to create more interesting surveys that hold respondents' attention better. Our results suggest that their can be a down side to this capability—the very same features that attract the respondents' attention may sometimes change their answers.

5. References

Aquilino, W., & LoSciuto, L. (1990) Effect of interview mode on self-reported drug use. *Public Opinion Quarterly*, *54*, 362-395.

Beebe, T. J., Harrison, P.A., McRae, J.A., Anderson, R.E., and Fulkerson, J.A. (1998). An evaluation of computer-assisted self-interviews in a school setting. *Public Opinion Quarterly, 11*, 623-632.

Boekeloo, B., Schiavo, L., Rabin, D., Conlon, R., Jordan, C., & Mundt, D. (1994). Self-reports of HIV risk factors at a sexually transmitted disease clinic: Audio vs written questionnaires. *American Journal of Public Health*, *84*, 754-760.

Couper, M.P. (1998), Review of Byron Reeves and Clifford Nass, 'The Media Equation: How people treat computers, television, and new media like real people and places'. *Journal of Official Statistics*, 13 (4): 441-443.

Crowne, D., & Marlowe, D. (1964). *The approval motive.* New York: John Wiley.

Fogg, B.J., and Nass, C. (1997). Silicon sycophants: The effects of computers that flatter. *International Journal of Human-Computer Studies*, *46*, 551-561.

Hatchett, S., and Schuman, H. (1975-76). White respondents and race-of-interviewer effects. *Public Opinion Quarterly, 39*, 523-528.

Hochstim, J. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, *62*, 976-989.

Kane, E.W., and Macauley, L.J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly, 57*, 1-28.

Lessler, J. T., and O'Reilly, J. M. (1997). Mode of interview and reporting of sensitive issues: Design and implementation of audio computer-assisted self-interviewing. In L. Harrison & A. Hughes (Eds.), *The validity of self-reported drug use: Improving the accuracy of survey estimates* (pp. 366-382). Rockville, MD: National Institute on Drug Abuse.

London, K., & Williams, L. (1990, May). A comparison of abortion underreporting in an in-person interview and self-administered questionnaire. Paper presented at the Annual Meeting of the Population Association of America, Toronto, Canada.

Kiesler, S., and Sproull, L. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50*, 402-413.

Mosher, W. D., & Duffer, A. P., Jr. (1994, May). *Experiments in survey data collection: The National Survey of Family Growth Pretest.* Presented at the meeting of the Population Association of America, Miami.

Moon, Y. (1998). Impression management in computer-based interviews: The effects of input modality, output modality, and distance. *Public Opinion Quarterly, 62*, 610-622.

Mott, F. (1985). *Evaluation of fertility data and preliminary analytic results from the 1983 Survey of the National Longitudinal Surveys of Work Experience of Youth.* Report to the National Institute of Child Health and Human Development by the Center for Human Resources Research, January, 1985.

Nass, C., Fogg, B.J., and Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45*, 669-678.

Nass, C., Moon, Y., and Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology, 27*, 864-876.

Paulhus, D.L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609.

Presser, S., and Stinson, L. (1998). Data collection mode and social desirability bias in self-reported religious attendance. *American Sociological Review, 63*, 137-145.

Reeves, B., and Nass, C. (1997). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge: CSLI and Cambridge University Press.

Schuman, H., and Converse, J. (1971). The effects of black and white Interviewers on white respondents in 1968. *Public Opinion Quarterly, 35*, 44-68.

Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., and Water, K. (1996), "When the Interface Is a Face." *Human-Computer Interaction*, 11: 97-124.

Tourangeau, R., Rasinski, K., Jobe, J.B., Smith, T.W., and Pratt, W.F. (1997). Sources of error in a survey on sexual behavior. *Journal of Official Statistics, 13*, 341-365.

Tourangeau, R., and Smith, T. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*, 275-304.

Tourangeau, R., and Smith, T. (1998). Collecting sensitive information with different modes of data collection. In M. Couper, R. Baker, J. Bethlehem, C. Clark, J. Martin, W. Nicholls, and J. O'Reilly, (Eds.), *Computer assisted survey information collection.* New York: Wiley.

Turner, C.F., Ku, L., Rogers, S.M., Lindberg, L.D., Pleck, J.H., and Sonenstein, F.L. (1998). Adolescent sexual behavior, drug use and violence: Increased reporting with computer survey technology." *Science, 280*, 867-873.

Walker, J.H., Sproull, L., and Subramani, M. (1994). Using a human face in an interface. *Proceedings of the Conference on Human Factors in Computers '94*, 85-91. Boston: ACM.

Acknowledgments

The research reported here was supported by a grant from the National Science Foundation (SES-9907395), awarded to Roger Tourangeau, Mick Couper, and Bob Tortora. We gratefully acknowledge NSF's support. We also thank Margrethe Montgomery and Bill Sukstorf of The Gallup Organization for their invaluable assistance in carrying out these studies.