Title
Clarifying Word Meanings in Computer-Administered Survey Interviews

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 22(22)

Authors
Schober, Michael F.
Conrad, Frederick G.
Bloom, Jonathan E.

Publication Date
2000

Peer reviewed

Clarifying Word Meanings in Computer-Administered Survey Interviews

Michael F. Schober (schober@newschool.edu)
New School for Social Research; Department of Psychology, AL-340; 65 Fifth Avenue
New York, NY 10003 USA

Frederick G. Conrad (conrad_f@bls.gov)
Room 4915; Bureau of Labor Statistics; 2 Massachusetts Ave. NE
Washington, DC 20212 USA

Jonathan E. Bloom (jonathan_bloom@dragonsys.com)
Dragon Systems, Inc.; 320 Nevada St.; Newton, MA 02160 USA

Abstract

We investigated the extent to which a collaborative view of human conversation transfers to interaction with non-human agents. In two experiments we contrasted user-initiated and mixed-initiative clarification in computer-administered surveys. In the first study, users who could clarify the interpretations of questions by clicking on highlighted text comprehended questions more accurately (in ways that more closely fit the survey designers' intentions) than users who couldn't, and thus they provided more accurate responses. They were far more likely to obtain help when they had been instructed that clarification would be essential than when they were merely told it was available. In the second study, users interacting with a simulated speech interface responded more accurately, and asked more questions, when they received unsolicited clarification about question meaning from the system in response to their linguistic cues of uncertainty. The results suggest that clarification in collaborative systems will only be successful if users recognize that their own conceptions may differ from the system's, and if they are willing to take extra turns to improve their understanding.

Introduction

Saying something doesn't guarantee it will be understood. People engage in dialog to make sure that what the speaker intended has been understood—to ground their understanding (e.g., Clark & Brennan, 1991; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). People ground their understanding to a criterion sufficient for their current purposes; in casual conversations (e.g., at a cocktail party), people may not need to understand precise details to satisfy their conversational goals, but in other settings (e.g., air traffic control tower conversations, calls to a technical help desk when your computer crashes, or conversations with your ex-spouse about child visitation) the stakes are higher.

This collaborative view of human conversation differs from traditional accounts of language use (what Akmajian et al., 1990 called the "message model" of communication), where listeners interpret utterances directly. The traditional view is that the meaning of an utterance is contained within the words themselves, and that the process of comprehension involves looking up those meanings in the mental dictionary and combining them appropriately; a collaborative view argues that accurate comprehension also requires dialog so that people can clarify what is meant (see Clark, 1996).

In the studies reported here we investigate the extent to which this collaborative view of human conversation transfers to interaction with non-human agents, and we examine whether a collaborative view can improve user interface design. Examining collaboration in human-computer interaction forces us to specify details of the collaborative view that can test its limits and refine our theories of human communication.

We contrast two approaches to designing collaborative systems that support the clarification of word meanings. Under one approach, clarification is user-initiated—that is, if the user explicitly requests clarification, the system provides it. This requires users to recognize that they need clarification and to be willing to ask for it. Under the other approach, clarification is mixed-initiative—that is the system also provides (or offers to provide) clarification when it diagnoses misunderstanding, based on user behavior. For example, in a desktop or speech interface a system could provide clarification when the user takes too long to act; in a speech interface a system could provide clarification when the user's speech is hesitant or disfluent (containing *um*s and *uh*s, restarts, etc.).

We examine these issues in the context of survey interviewing systems, where systems present questions and users answer them. To our knowledge, current dialog systems for surveys (see Couper et al., 1998 on "computerized self-administered questionnaires") do not allow either user-initiated or mixed-initiative clarification of meaning. Rather, they embody strict principles of standardization developed for human-human interviews, where the interpretation of questions should be left entirely up to respondents (e.g., Fowler & Mangione, 1990). The argument for standardization is that if interviewers help respondents to interpret questions, they might influence responses, but if interviewers read scripted questions and provide only "neutral" feedback, responses are less likely to be biased. We have demonstrated that in human-human interviews even supposedly nonbiasing feedback by interviewers can affect responses (Schober & Conrad, 1997, in press). More importantly, strict standardization can actually harm data quality because it prevents respondents from grounding their understanding of the questions. This is a problem because people's interpretations of seemingly straightforward questions like "How many bedrooms are there in your house?" can vary enormously; without grounding their understanding of questions, respondents may conceive of questions in unintended ways, and the resulting data may not fulfill the survey designers' purposes (Clark & Schober, 1991). We have shown that responses in strictly standardized interviews can be less accurate than responses in more interactive interviews where respondents can ground their understanding of questions with the interviewers (Conrad & Schober, 2000; Schober & Conrad, 1997).

The task of responding to a computerized survey differs from many human-computer interaction situations. First, in survey systems users provide information to the system rather than retrieving information from the system, as with a database query system or a web search interface. Second, survey system users' need for precise understanding may be lower than when they interact with other systems. Users may care less about precisely understanding the words in survey questions when providing opinions to market researchers (misunderstanding has few consequences for the user) than understanding the words in an on-line job application or an on-line health claims form (where misunderstandings can be costly).

Experimental Methods

In our studies we assess whether systems that enable users to clarify the survey concepts do actually lead to improved comprehension of questions (and thus improved response accuracy), as a collaborative theory would predict. We examine the effects of clarification on task duration—clarification probably takes more time, and this may offset any benefits. We also examine the effects of clarification on user satisfaction; even if clarification improves comprehension, it could be annoying.

Our first study (Conrad & Schober, 1999) uses a desktop interface, in which the computer displays questions on a screen. The user enters responses and asks for clarification with the keyboard and mouse. Our second study (Bloom, 1999) uses an interface, in which questions are presented in a synthesized voice through a headset. The user answers questions and asks for clarification by speaking into the headset microphone.

In both studies, all users were asked the same survey questions, which had been used in earlier studies of human-human survey interviews (e.g. Schober & Conrad, 1997). We adapted 12 questions from three ongoing U.S. government surveys. Four questions were about employment, from the Current Population Survey (e.g., "Last week, did you do any work for pay?"); four questions were about housing, from the Consumer Price Index Housing survey (e.g., "How many people live in this house?"); four questions were about purchases, from the Current Point of Purchase Survey (e.g., "During the past year, have you purchased or had expenses for household furniture?"). For each question, the survey designers had developed official definitions for the key concepts, which clarified whether, for example, a floor lamp should be considered a piece of household furniture, or whether a student away at college should be considered to be living at home.

Users answered these questions on the basis of fictional scenarios, so that we could measure response accuracy—that is, the fit between users' answers and the survey designers' official definitions. For each question there were two alternate scenarios, one typical and one atypical. With the typical scenario, the survey question was designed to be easy for users to interpret—to map onto the user's (fictional) circumstances in a straightforward way. For example, for the question "Has Kelley purchased or had expenses for household furniture?", the typical scenario was a receipt for an end table, which is clearly a piece of furniture. With the atypical scenario, it was less clear how the survey question should be answered. For example, for the household furniture question the atypical scenario was a receipt for a floor lamp, which is harder to classify without knowing the official definition of "household furniture."

For each user, half the scenarios described typical situations and half atypical situations.

Study 1: Desktop interface

In this study, we varied the way the survey system provided clarification. When clarification was user-initiated, users could request the official definition for a survey concept by clicking the mouse on highlighted text in the question. When clarification was mixed-initiative, the system would also offer a definition when users were "slow" to respond. This was defined as taking longer than the median response time for atypical scenarios when no clarification was available. This offer was presented as a Windows dialog box; users could reject the offer by clicking "no" if they didn't think clarification was needed.

We also varied instructions to the users about how precisely they would need to understand the system's questions—that is, we varied the grounding criterion. Some users were told that clarification was essential; they were encouraged to obtain definitions from the computer because their everyday definitions might differ from the survey's. Other users were told merely that clarification was available, that definitions would be available if users wanted them. The five experimental conditions are displayed in Table 1.

54 users, recruited from an advertisement in the Washington Post, were paid to participate. Most (44) reported using a computer every day.

Table 1: Experimental conditions, Study 1.

Type of clarification	User instructed that…
1 no clarification	
2 at user's request	Clarification essential
3 at user's request	Clarification available
4 when user is slow or at user's request	Clarification essential
5 when user is slow or at user's request	Clarification available

Results

Users' responses were almost perfectly accurate (their responses fit the official definitions) when they answered about typical scenarios. For atypical scenarios, users were more accurate when they could get clarification than when they couldn't (see Figure 1). Response accuracy mainly depended on the instructions to the user about the grounding criterion. When users had been told that definitions were merely available, their accuracy was as poor as when they couldn't get clarification. When they had been told that definitions were essential, response accuracy was much better, whether users had to request clarification, $F(1,49) = 9.82$, $p < .01$, or whether the system also offered it, $F(1,49) = 14.38$, $p < .01$.

As Figure 2 shows, response accuracy was strongly related to how often users received clarification. When users had been told that definitions were essential, they requested clarification most of the time; in fact, they frequently requested it for typical scenarios, when presumably it wasn't necessary. They also requested clarification quickly, which meant that when the system could also provide clarification (conditions 4 and 5) it rarely did. In contrast, users who had been told that clarification was merely available rarely asked for it, and they responded to the questions so quickly that system-initiated clarification was rarely triggered. Apparently, it didn't occur to these users that their interpretation of ordinary terms like "bedroom" and "job" might differ from the system's, and so they answered confidently, quickly, and inaccurately.

As Figure 3 shows, clarification took time. Response times were much longer in cases where users received clarification. As we anticipated, improved accuracy from clarification can be costly.

Users' ratings of their satisfaction with the system suggested two things. First, users who could not get clarification reported that they would have asked for clarification if they could. This suggests that interacting with dialog survey systems that don't allow clarification may be relatively unsatisfying. Second, users' grounding criteria affected their perceptions of the system. System-initiated clarification was rated on a 7 point scale as useful (6.0) and not annoying (1.0) by "clarification essential" users, and less useful (3.9) and more annoying (4.25) by "clarification available" users. Presumably users who had been told that clarification was available found it jarring for the system to offer unsolicited help for seemingly straightforward questions.

Overall, these results suggest that the success of human-machine collaboration may depend both on users' grounding criteria—how important they believe it is to understand accurately—and also on whether users recognize that system concepts may differ from theirs.

Study 2: Speech interface

This study used a Wizard-of-Oz technique to simulate a speech interface. Users believed they were interacting with a computer, when actually a hidden experimenter presented the questions and scripted clarification. To enhance believability, we used an artificial-sounding computer voice (Apple's "Agnes" voice).

This study used exactly the same questions and scenarios as Study 1. Users participated in one of four experimental conditions. In the first condition, the system never provided clarification. In the second condition, clarification was user-initiated—the system would provide clarification if users asked for it explicitly. In the third condition, the initiative was mixed—the system would "automatically" provide full definitions when users displayed specific uncertainty markers that had been shown to be more prevalent in atypical situations in human-human interviews collected with these materials (Bloom & Schober, 1999). These included *um*s, *uh*s, pauses, repairs, and talk other than an answer. In the fourth condition, the system always provided clarification; no matter what the user did, the system would present the full official definition for every question.

40 users recruited from an advertisement in the *Village Voice* were paid to participate.

Results

As in Study 1, users' responses were almost perfectly accurate when they answered about typical scenarios. For atypical scenarios, users were substantially more accurate when they were always given clarification (80%) than when they were never given clarification (33%), $F(1,36) = 10$, $p < .005$. When users initiated clarification, their response accuracy was no better (29%) than when they were never given clarification, because they almost never asked for it. As in Study 1, it seems likely that it didn't occur to users that clarification was necessary. Response accuracy was better when the initiative for clarification was mixed (59%), $F(1,36) = 10.11$, $p < .005$, although it was not as good as when clarification was given always.

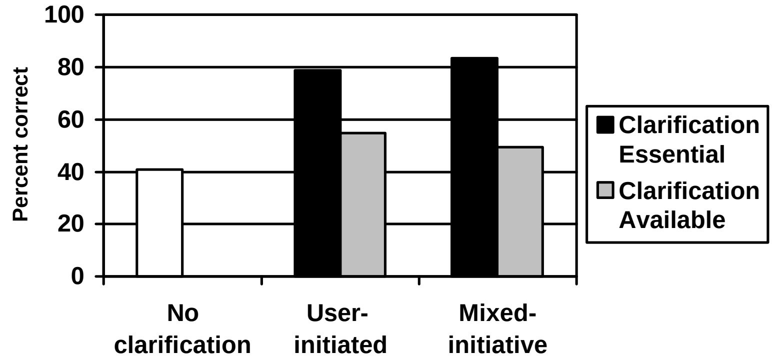

Figure 1: Response accuracy for atypical scenarios, Study 1

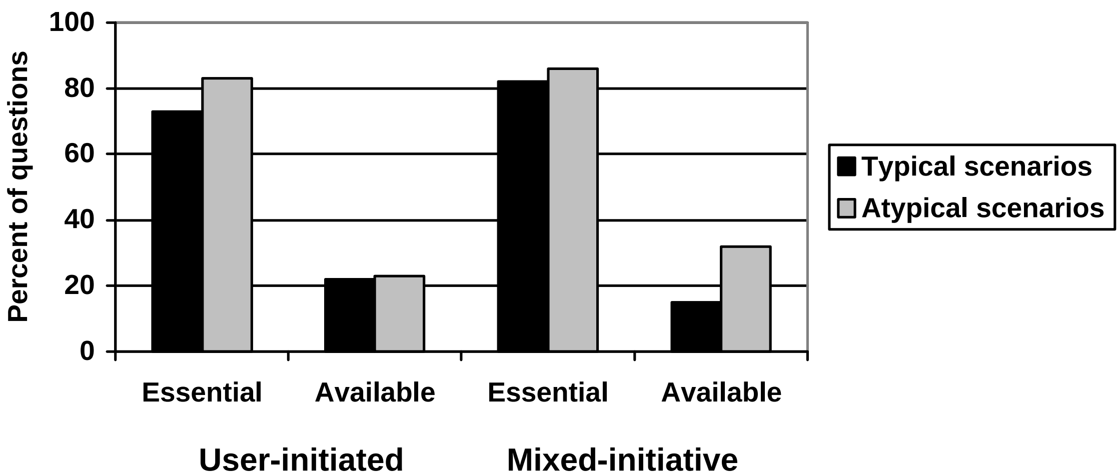

Figure 2: How often users received clarification, Study 1

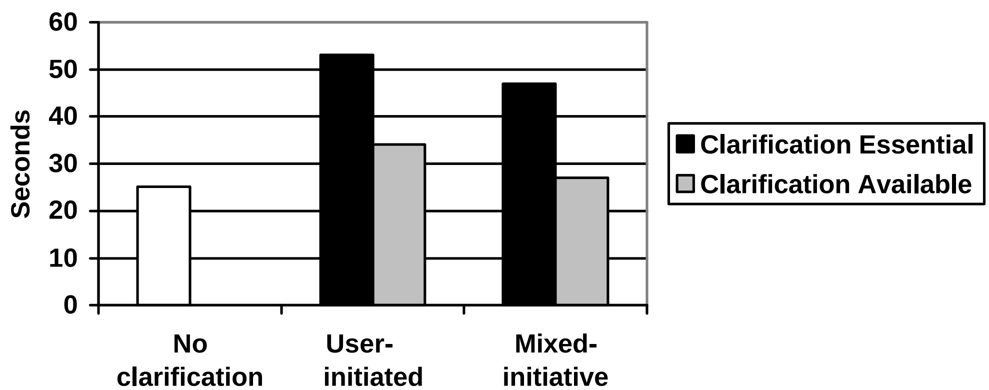

Figure 3: Response time per question, Study 1

System-initiated clarification increased the amount of user-initiated clarification: users were more likely to ask questions in the mixed-initiated condition, presumably because they were more likely to recognize that clarification might be useful. These users also spoke less fluently, producing more *um*s and *uh*s. We speculate that this was because these users at some level recognized that the system was sensitive to their cues of uncertainty.

Overall, the users in this study requested clarification far less often than the users in Study 1. This might result from any or all of the differences between our desktop and speech interfaces. In the speech interface, clarification was harder to request; requests had to be formulated into explicit questions rather than being accomplished by simple mouse clicks. Also, in the speech interface the definition unfolded over time (sometimes a substantial amount of time, up to 108 seconds), rather than appearing all at once, and in our application it was impossible to shut off; in the desktop interface, the definition appeared all at once and could be closed with a simple mouse click. Also, unlike in the desktop study, users couldn't reject system-initiated offers of clarification; here the system immediately provided clarification when triggered, without giving the option of rejecting the help.

As in Study 1, clarification took time. The more clarification a user received, the more time the interviews took. Sessions where clarification was always provided took more than twice as long as sessions with no clarification or when it was (rarely) user-initiated (12.8 versus 5.2 and 4.9 seconds per question, respectively); mixed-initiative clarification took an intermediate amount of time (9.6 seconds per question).

Also as in Study 1, users rated the system more positively when it was responsive (user- or mixed-initiative conditions). When the system was not responsive (no clarification or clarification always), users wanted more control and felt that interacting with the system was unnatural. Users didn't report finding system-initiated clarification particularly more annoying than user-initiated clarification—which they almost never used.

Overall, these results suggests that enhancing the collaborative repertoire of a speech system can improve comprehension accuracy without harming user satisfaction, as long as the system provides help only when it is necessary. But these improvements come at the cost of increased task duration, which could make such systems impractical in real-world survey situations.

Conclusions

Our findings demonstrate that a collaborative view can indeed transfer to interaction with non-human agents. Increased system clarification abilities can improve users' comprehension (and thus their response accuracy), while increasing (or not reducing) user satisfaction. But this comes at the cost of increased task duration, which could lower survey completion rates in the real world.

Our findings also demonstrate that extended clarification sequences are likely to be rare or unnecessary when users' conceptions are likely to be the same as the system's, as in our typical scenarios. The need for building survey systems with enhanced collaborative abilities may depend on the likelihood of potential misunderstandings; if this likelihood is high or unknown, enhanced collaborative abilities may be worth implementing.

The benefits we have shown for collaboratively enhanced survey systems come even with our rudimentary implementations, which are based on the most generic of user models (see Kay, 1995). A stronger test of collaborative approaches requires more customized interfaces, in which, for example, the system would reason about which parts of definitions would be appropriate to present at any given moment, what particular users are likely to misunderstand, etc. (see Moore, 1995).

Our findings demonstrate that computer implementations of surveys seem to run into exactly the same problems as human-human survey and instructional situations, where people don't always recognize they need help or aren't willing or able to ask for help (e.g., Graesser & McMahen, 1993; Schober & Conrad, 1997).

But our findings also show that in some situations (our desktop interface, when users were told that clarification was essential), users are indeed willing to ask for clarification more often than they are with human interviewers (Schober & Conrad, 1997). This is consistent with findings in other domains that interaction with a computer can lead to better task outcomes than interaction with a person. For example, people may be more willing to accept correction from an intelligent computer tutor than from a human tutor (Schofield, 1995), and people are more willing to admit to undesirable behaviors when asked about them on self-administered computer surveys than in human-administered surveys (Tourangeau & Smith, 1996).

We propose that some of these improvements from interacting with computers don't arise simply from the fact that the computer isn't a person. They arise in part from the fact that the costs and constraints of grounding vary in different media, as Clark and Brennan (1991) argued. Most tutoring and survey systems to date have been direct manipulation or simple (textual) character entry systems like our desktop interface; in such interfaces the user's costs of requesting information from the system can be low. The human interactions to which such systems are often compared are speech interactions, where people have to formulate clarification requests explicitly and clarification takes significant amounts of time. Any differences in task performance may just as likely result from the differences between direct manipulation and speech as from the differences between computers and humans.

We believe our findings also require us to refine a theory of human-human collaboration by explicitly introducing the notion of initiative. Our findings that comprehension success can vary depending on whether the user or system takes the initiative should be extended to the human realm; a collaborative theory should include who takes the responsibility for clarifying meaning. In many cases speakers are responsible for what they mean, and listeners assume that what speakers say is readily interpretable to them in the current context (the "interpretability presumption," in Clark and Schober's [1991] terms). But in situations where the speaker is less competent than the addressee, the addressee may take responsibility for the meaning, and may initiate clarification (Schober, 1998). Who should be responsible under what circumstances, and what determines how speakers decide whose effort should be minimized, are important questions for a theory of collaboration.

Altogether, our results suggest that user-initiated clarification will work only if users recognize that clarification will help, recognize that the system's concepts may differ from theirs, are motivated to understand precisely, and are willing to take the extra turns to ground understanding. Explicit instructions to users can help make this happen—help set a high grounding criterion—but it's unclear whether such instruction is feasible in real-world situations. Our results suggest that system-initiated clarification will work only if users give reliable evidence of misunderstanding and if they are willing to accept offers of clarification. It won't work if users are confident in their misinterpretations.

In general, the opportunity for clarification dialog won't help if users don't recognize it's needed.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No SBR9730140 and by the Bureau of Labor Statistics. The opinions expressed here are those of the authors and not of the Bureau of Labor Statistics. We thank Susan Brennan and Albert Corbett for help with an earlier version of this paper.

References

Akmajian, A., Demers, R.A., Farmer, A.K., & Harnish, R.M. (1990). *Linguistics: An introduction to language and communication, 3rd ed.* Cambridge, MA: MIT.

Bloom, J.E. (1999). *Linguistic markers of respondent uncertainty during computer-administered survey interviews*. Doctoral dissertation, Department of Psychology, New School for Social Research, New York.

Bloom, J.E., & Schober, M.F. (1999). Respondent cues that survey questions are in danger of being misunderstood. In *Proc. of the ASA*, *Section on Survey Research Methods* (pp. 992-997). Alexandria, VA: American Statistical Association.

Clark, H.H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. In L.B. Resnick, J.M. Levine, & S.D Teasley, (Eds.), *Perspectives on socially shared cognition*. Washington, DC: APA.

Clark, H.H., & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive Science*, *13*, 259-294.

Clark, H.H., & Schober, M.F. (1991). Asking questions and influencing answers. In J.M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys*. New York: Russell Sage Foundation.

Clark, H.H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22,* 1-39.

Conrad, F.G., & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. *Proc. of the 3rd ASC Int'l Conference* (pp. 91-102). Chesham, UK: Association for Survey Computing.

Conrad, F.G., & Schober, M.F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64,* 1-28.

Couper, M.P., et al. (Eds.) (1998). *Computer assisted survey information collection.* New York: Wiley.

Fowler, F.J., & Mangione, T.W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: SAGE Publications, Inc.

Graesser, A.C., & McMahen, C.L. (1993). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. *Journal of Educational Psychology*, *85*, 136-151.

Kay, J. (1995). Vive la difference! Individualized interaction with users. In C.S. Mellish, (Ed.), *Proc. of the 14th International Joint Conference on Artificial Intelligence,* 978-984. San Mateo, CA: Morgan Kaufmann.

Moore, J.D. (1995). *Participating in explanatory dialogues: Interpreting and responding to questions in context.* Cambridge, MA: MIT.

Schober, M.F. (1998). Different kinds of conversational perspective-taking. In S.R Fussell & R.J. Kreuz (Eds.), *Social and cognitive psychological approaches to interpersonal communication*. Mahwah, NJ: Erlbaum.

Schober, M.F., & Clark, H.H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*, 211-232.

Schober, M.F., & Conrad, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, *61*, 576-602.

Schober, M. F., and. Conrad, F. G. (In press). A collaborative view of standardized survey interviews. In D. Maynard, H. Houtkoop, N. C. Schaeffer, & J van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview,* New York: Wiley.

Schofield, J.W. (1995). *Computer and Classroom Culture*. Cambridge, UK: Cambridge University Press.

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