CHAPTER 8

New Frontiers in Standardized Survey Interviewing

Frederick G. Conrad
Michael F. Schober

Sample surveys are the empirical backbone of the social sciences, government policy, political campaigns, and corporate strategy. The method of collecting data that is generally assumed to provide the highest quality information is the standardized interview, either on the telephone or face-to-face. There are exceptions, of course, such as the collection of data about sensitive topics like drug use and sexual activities, for which self-administration (i.e., without an interviewer) is believed to increase the honesty of respondents' answers. However, by and large, if social researchers can afford to conduct interviews, they do so. Interviews generally lead to higher response rates than self-administration (particularly when conducted face-to-face) and allow respondents who cannot read (including those who are visually impaired) to participate.

Standardized interviewing is an approach to collecting survey data in which interviewers read questions exactly as worded to every respondent and are trained never to provide information beyond what is scripted in the questionnaire. The goal is to increase comparability across interviews so that different answers from different respondents cannot be attributed to different question stimuli. In principle, standardized interviewers are interchangeable; the answers should be the same no matter who asks the question.[1]

Uniform question presentation has not always been the industry standard (see Beatty, 1995, for a historical perspective on standardized interviewing). However, since the widespread deployment of telephone interviewing in the 1970s, standardization has been virtually synonymous with scientific social research (see Fowler & Mangione, 1990, for a clear statement of standardized interviewing practice and goals). Nonetheless, standardized interviewing has been criticized because the limits it places on what interviewers can say may prevent them from ensuring that questions are understood as intended. Thus standardized interviews might produce reliable results—that is, the same question produces the same answer, irrespective of who asks it—without producing valid results, because respondents may systematically misunderstand the question, effectively answering a question other than what the researchers intended to ask (Suchman & Jordan, 1990).

Our work (e.g., Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004) has focused on the inaccuracies in survey responses that can arise from interviewers' inability to "ground" the meaning of questions (e.g., Clark & Schaeffer, 1989). The research is motivated by the fact that, in ordinary discourse, speaker and listener can achieve mutual understanding by discussing the meaning of what has been said; but in standardized survey interviews, the interviewer may not clarify what has been said if to do so requires more than a verbatim repetition of the question. It stands to reason that response accuracy should be inferior in just those cases in which respondents find questions to be ambiguous, for example, in which it is not clear which behaviors and events respondents should include and which they should exclude. The question for us is what the consequences are of licensing interviewers in such situations to say whatever they judge necessary to resolve the ambiguity and clarify the intended meaning of the question. We have called this approach "conversational interviewing" because interviewers are able to ground question meaning as in ordinary conversation. Conversational interviewing is similar in spirit to standardized interviewing—the goal being to produce comparable data—but emphasizes uniform *interpretation* rather than uniform *wording*.

In the first part of this chapter we review some of our research comparing strictly standardized interviewing with variants of conversational interviewing. We have observed substantial improvement in response accuracy when interviewers can help respondents determine which behaviors they should include and which they should exclude. Throughout, we focus on response accuracy rather than on more traditional survey measures of data quality such as response variance or missing data.

In the second part of the chapter we consider the implications of bringing ordinary conversational practices such as clarifying what has been said into automated self-administered survey "interviews." With new technologies, survey designers have the opportunity to explore a range of data collection methods in which computer interfaces incorporate more and more features of human interviewing. Text-based Web surveys can now add recorded human or synthesized voices, video (either live or recorded), or animated faces with varying degrees of verisimilitude and varying degrees of dialogue capability. We first examine the impact on data quality and user satisfaction of building clarification dialogue into textual questionnaires displayed in Web browsers and speech dialogue interviewing systems. We then consider some of the issues created when virtual interviewers—that is, avatars or animated agents—ask questions and clarify their meanings. In general, the issue concerns the degree to which respondents treat the interviewing agents as human-like or computer-like, but the issue is slightly different when the interviewing agents ask sensitive versus nonsensitive questions. Because designers can control so many features of the interviewing agent (e.g., facial appearance and vocal characteristics) and the way it interacts with respondents (e.g., its ability to engage in human-like dialogue), methodologists are forced to reconsider what standardization and comparability mean. For example, it is already possible to match features of the interviewing agent—gender, ethnicity, or age of voice or appearance—to those of each respondent. But when is this desirable and when is it undesirable?

Theories of Communication and Survey Interviews

Most large-scale surveys try to standardize what interviewers say. As described by Fowler and Mangione (1990), interviewers must read questions exactly as worded, they must probe neutrally, and they must never allow their own ideas to influence the respondents' answers. According to Fowler and Mangione, this practice reduces or eliminates what they call *interviewer-related error*—any systematic effect of particular interviewers on survey responses.

This view grows out of a long tradition in survey methodology of distinguishing between different sources of error and using different methods for reducing each type of error. For example, this view holds that respondent comprehension error is best handled by wording the question in ways that most respondents are likely to understand. By pretesting early versions of questions and revising those questions on the basis of the pretest results, misunderstanding can be largely reduced. Standardized interviewing presupposes that questions have been pretested and are universally interpretable. The argument is that if interviewers deliver pretested questions in a standardized way—exposing respondents to the same question stimulus (Fowler & Mangione, 1990, p. 14)—then researchers can be confident that differences in the answers stem only from the actual differences between respondents and not from any misunderstanding by some respondents or idiosyncrasies of interviewers' behavior (p. 15).

However, by separating interviewer behavior, respondent behavior, and question wording, this prevailing approach relies on a view of communication that has been discredited, at least for ordinary spontaneous conversations. This view, dating back to John Locke or perhaps even earlier, has been called, variously, the *message model* (Akmajian, Demers, Farmer, & Harnish, 1990), the *conduit metaphor* (Reddy, 1979/1993), and the *meaning-in-words assumption* (Schober, 1998). According to this view, speakers encode their thoughts into linguistic messages and send these messages to recipients by speaking, who decode them into their own thoughts. Thoughts and conceptual material are thus transferred from one head to the other via words.

Although the message model captures most people's intuitions about how communication works, it cannot account for all of what goes on in ordinary conversations (see also Clark, 1992, 1996; Gibbs; 1994; Maynard & Whalen, 1995). The problem is that the message model assumes that the meaning of speakers' words is in the words themselves. But it is not. Rather, speakers can use the same words to express vastly different meanings on different occasions (see Akmajian et al., 1990). Consider the word *red* (Clark, 1996). It denotes one color when referring to wine, another when referring to hair, another when describing a fire truck, and so on. Suessbrick, Schober, and Conrad (2000) observed that when respondents answered the question "Have you smoked at least 100 cigarettes in your entire life?" 54% of listeners interpreted this as including "cigarettes that you took a puff or two from," 23% interpreted this as referring to "cigarettes you partially smoked," and 23% interpreted it as meaning "only cigarettes that you finished." Given this sort of variability in the meaning of words, how do people ever manage to communicate?

One proposal is Clark and Wilkes-Gibbs's (1986) *collaborative* model (see also Clark, 1992, 1996; Clark & Brennan, 1991; Clark & Schaefer, 1987, 1989; Schober & Clark, 1989). This model generalizes observations by Paul Grice, Emanuel Schegloff, and others, bringing them into the psychological realm in ways that can be modeled and tested precisely (e.g., Cahn & Brennan, 1999). Under the assumptions of this approach, as people speak, they carefully monitor their addressees for evidence of understanding or misunderstanding, and they adjust their utterances, moment by moment, to ensure that their addressees understand them well enough for current purposes. Addressees, by providing such evidence, help mold the utterances speakers produce.

In this view, no utterance is complete until it has been *grounded*—until both participants have accepted that it has been understood. Understanding a reference in any particular utterance requires active participation by both speaker and addressee, and this can take several turns. Note that the point is not that words do not have conventional meanings; in fact, the conventional meanings of words provide important constraints on speakers' meanings. But speakers regularly use words in idiosyncratic ways that go far beyond dictionary definitions (see Clark, 1991; Clark & Gerrig, 1983). Speakers produce utterances based on their *common ground* with their conversational partners—that is, what they presume that they and their conversational partners mutually know, believe, and assume.

Experimenting with Alternatives to Standardized Interviewing

What might the collaborative view of language use imply for survey interviews? As we indicated earlier, standardized interviewers are prohibited from grounding the meaning of questions (although the rationale is not expressed in these collaborative terms). If a respondent asks a standardized interviewer to clarify a question (e.g., "What do you mean by 'household'?") the interviewer can reply only by administering one of a small set of "neutral" or nondirective probes, such as rereading the question or indicating that the interpretation of the question is the sole obligation of the respondent (e.g., "Whatever it means to you"). By withholding substantive clarification (as in rereading the question) and even encouraging respondents to interpret the question in their own way, the interviewer relinquishes control over what the question means to the respondent. The obvious cost of this practice is that respondents may not understand the questions the way the survey researchers intended and thus answer them inaccurately, jeopardizing the validity of the research.

We have conducted a series of experiments in which we have evaluated response accuracy and interaction patterns for strictly standardized and more collaborative "conversational" interviews, as well as for a range of interviewing techniques in between. In these studies, professional interviewers telephoned naive respondents either in the laboratory (Schober & Conrad, 1997; Schober et al., 2004) or at home (Conrad & Schober, 2000) and asked questions from large U.S. government surveys. In the laboratory studies, the "respondents" answered the questions on the basis of scenarios that described the work, housing, and purchases of fictional people. Because we created the scenarios, we knew the correct answers—according to official definitions—for each question–scenario combination, and so we could determine response accuracy. In the household study, the respondents answered questions about their own lives; we evaluated the response accuracy by less direct reinterview methods.

In each study, respondents participated in either strictly standardized interviews, following Fowler and Mangione's (1990) prescriptions, or less standardized, more collaborative interviews. In all interviews, questions were first posed exactly as worded, underscoring the shared commitment to standardization under the two approaches; the difference just concerns whether it is wording or underlying meaning that is standardized. In the more collaborative interviews, interviewers were then encouraged to ground understanding of question meaning, for example, by providing scripted definitions when respondents explicitly asked for them (Schober et al., 2004) or by using whatever words interviewers chose to make sure respondents understood the questions as intended (Schober & Conrad, 1997).

In all studies our basic question is which interviewing technique leads to more accurate responses. Of course, even in purely standardized surveys, interviewers can affect responses. What we have tested here is how the kinds of influences that occur in strictly standardized interviews affect response accuracy as compared with the kinds of influences that occur in more collaborative interviews.

Mapping Ambiguity

In all studies we used pretested questions from ongoing surveys whose words and grammar have been shown to be understandable but whose interpretation for some respondents on some occasions might be unclear. Consider a question such as "During the past year, have you purchased or had expenses for household furniture?" A respondent who has bought an end table should not have much trouble answering "yes," but a respondent who has bought a floor lamp may be less sure. Or consider a question such as "Last week, how many hours did you work?" This should be clear for a respondent who has a 9-to-5 job that includes overtime but less clear for a respondent who does business over lunch or solves work-related problems while jogging. We presume that the second respondent would be more likely than the first to request clarification, such as, "What do you mean by work?" or "Should I consider business lunches to be work?" We refer to this as a mapping ambiguity: the mapping between the question term and the respondent's circumstances is unclear. Interviewers following strictly standardized procedures cannot meaningfully resolve a mapping ambiguity; they would be obliged to use neutral probing techniques, including "whatever it means to you." More conversationally flexible interviewers could provide information to the respondent that would help clarify what the author of the question had in mind.

In our laboratory studies, we designed the fictional scenarios so that half corresponded to the questions in a straightforward way (straightforward mappings) and the other half corresponded to the questions in a more complicated way (complicated mappings). In the household study, we had no control over the frequency of complicated mappings. Our prediction was that response accuracy would be high for both standardized and more conversational interviewing when the mappings were straightforward; when the mappings were complicated, accuracy should suffer for strictly standardized interviewing but not as much—maybe not at all—for conversational interviewing. This pattern of results in the household study, in which mappings were not under our control, would indicate that complicated mappings are frequent enough in the real world to warrant further exploration of more collaborative interviewing techniques.

In order to ensure that conversational interviewers could answer respondents' substantive questions, we needed to teach them the official definitions of key concepts in the questions. Providing standardized interviewers with this knowledge might seem to violate the principles of standardization: The only role for definitions in the standardized interview is to be read in their entirety to all respondents on all occasions or not read at all. But the logic of our experiments required us to train all interviewers together on the concepts so that any accuracy differences could not somehow be attributed to different levels of knowledge between standardized and conversational interviewers. Standardized interviewers were told that the concept training was necessary so that interviewers would be able to judge when respondents had answered a question completely (see Beatty, 1995). In all studies the training lasted about 90 minutes; interviewers first studied the official definitions and then actively carried out exercises to ensure that they had grasped the concepts in detail.

After the concept training, interviewers were then trained in their respective interviewing techniques. The standardized instructions were based on guidelines that appeared in an interviewing manual for a survey on which many of these interviewers regularly worked and were consistent with Fowler and Mangione's (1990) approach. Using this material, we reviewed standardized question-asking and neutral probing techniques and supplemented this with role-playing exercises.

The interviewers who were trained to use more conversational techniques were instructed to initially read the question as worded. Then (depending on the study) they could substantively answer respondents' requests for clarification, either following a script or in their own words; they could also provide unsolicited clarification (scripted or in their own words) when respondents seemed to need it, even if respondents hadn't asked for help. In another study (Schober et al., 2004, Experiment 2), interviewers were not trained in a particular technique but were told to do whatever they ordinarily do.

Procedure

In the laboratory studies, respondents were given a packet of scenarios to study, and then they were questioned over the telephone about the scenarios. The respondents in conversational interviews were instructed to work with the interviewers to make sure they had interpreted the questions as the survey designers intended; they were encouraged to ask for clarification if they needed it. Response accuracy was measured as the percentage of questions for which responses matched what the official definitions indicated was correct.

Although the interviewers knew that respondents were answering on the basis of fictional situations, they were not familiar with the content of the individual scenarios, and so knowledge was allocated much as it is in real surveys: Interviewers knew the questions, and respondents knew about their own circumstances. We counterbalanced the items so that the respondents who were assigned to a particular interviewer always received different versions of the scenarios. This way the interviewers could not become familiar with the scenarios based on anything the earlier respondents might have said.

In the household study respondents were telephoned at home and asked about their own lives; no scenarios were involved. Because we could not directly determine the accuracy of the respondents' answers, we designed the experiment to provide two indirect measures. One measure was response change between interviews. All respondents participated in two interviews: The first was strictly standardized for all respondents, and the second was strictly standardized for half of the respondents and conversational for the other half. If respondents' circumstances mapped in a complicated way to the question concepts, they should be more likely to change their answers between an initial standardized interview and a subsequent conversational interview than between two standardized interviews. The reason is that in the conversational interviews the interviewers were instructed to clarify question meaning and correct respondent misconceptions, which could lead to different answers than provided in the initial interview. In contrast, in the second standardized interview, the interviewers were not permitted to clarify meaning or to correct initial respondent misconceptions; therefore, we hypothesized, responses would likely remain unchanged—reliable but incorrect.

The other measure in the household survey was the "legality" of respondents' explanations for their responses. If respondents answered "yes" when asked if they had made certain types of purchases, they were asked to briefly describe the purchase(s). These descriptions were then coded for their consistency with official definitions—their legality. For example, the definition of moving expenses explicitly excludes payments for do-it-yourself moving; a respondent who answered "yes" when asked if he or she incurred moving expenses and based this response on having rented a moving van would have provided an *illegal* explanation. When interviewers were licensed to clarify question meaning (in a conversational interview), we expected respondents to be more likely to base their responses on legal purchases than when interviewers were not able to clarify question meaning (in any of the standardized interviews).

Results

We first analyzed the transcripts of the interviews to verify that interviewers had followed our instructions and implemented the appropriate technique for each specific study. One way we demonstrated that interviewers followed instructions was by coding the various deviations from strict standardization and comparing their frequency between interviewing techniques. These deviations included rephrasing all or part of the question, providing all or part of a definition (either verbatim or paraphrased), converting the respondents' descriptions of their circumstances (the fictional scenario in the laboratory studies) into an answer, offering to provide clarification, confirming or disconfirming the respondent's interpretation of the question, and requesting particular information about the respondent's circumstances. For example, in the following exchange (from Schober & Conrad, 1997) the conversational interviewer paraphrased the long government definition of "household furniture" to answer the respondent's question:

> I: Has Kelly purchased or had expenses for household furniture?
>
> R: Um . . . is a lamp furniture?
>
> I: No sir, we do not include lamps and lighting fixtures.
>
> R: Okay, no.
>
> (I: goes on to next question)

In strictly standardized interviewing, the interviewer should not have answered the respondent's request for clarification, because by doing so she helped interpret the survey question for this respondent buy may not have done so for another. Across the various studies, our coding gives us confidence that our interviewer training led to fundamentally different types of interaction. In the Schober and Conrad (1997) study, for example, 85% of the question–answer sequences in conversational interviews contained deviations from standardization, compared with only 2% in strictly standardized interviews.

We can now turn to response accuracy. Again, an accurate response in our experiments is one that is consistent with the official definition of the relevant concept. In our earlier lamp example, the correct answer is "no," because a lamp purchase does not qualify as a furniture purchase for the purposes of the survey from which this question was drawn; therefore, the respondent's answer was accurate.

Across all our lab studies (Schober & Conrad, 1997; Schober et al., 2004, Experiments 1 and 2), a general pattern emerges. When mappings between question concepts and people's circumstances (scenarios) are straightforward, all interviewing techniques lead to nearly perfect accuracy; virtually all respondents interpret question concepts in the ways that survey designers intended. But when mappings between question concepts and people's circumstances are complicated, strictly standardized interviewing leads to quite poor response accuracy (28%). Response accuracy improves when interviewers can provide clarification on request (57%), and it improves substantially more when interviewers can both offer clarification that they believe respondents need as well as provide it when respondents request it (77%). Response accuracy is highest when the interviewers who can both volunteer clarification and provide it on request can also use their own words to clarify question concepts, even if this departs from what is scripted (87%). This last group of interviewers can exercise much of the discretion and flexibility that is typical of everyday conversation.

These results are mirrored in the household study (Conrad & Schober, 2000), in which we used indirect measures of response accuracy. Respondents changed their answers more often when their second interview was conversational (22%) than when it was strictly standardized (11%). In addition, more responses were based on legal purchases when the second interview was conversational (95%) than when it was standardized (57%). For example, in conversational interviews respondents who answered "yes" to the question about moving expenses were more likely to do so because they had hired commercial movers or had had other expenses included in the definition than in standardized interviews, in which they were more likely to incorrectly respond "yes" because of excluded expenses such as do-it-yourself moving. This result was primarily due to interviewers' explaining what to count and what not to count, that is, grounding the meaning of the term. The improved response accuracy in conversational interviews suggests that respondents' actual circumstances (as opposed to the fictional scenarios presented in the lab studies) are complicated sufficiently often—at least for these questions—to justify exploring the technique further.

Fowler and Mangione (1990) have raised the concern that interviewers whose wording is not strictly standardized will damage the quality of the responses. They are particularly concerned that in questions about opinions, nonstandardized interviewers may bias respondents by presenting their own opinions or by reacting to the respondents' answers. Regarding interviewers' explaining the intent of questions, Fowler and Mangione are concerned that nonstandardized interviewers will mislead respondents by providing inaccurate information. In our experiments this has not been the case. For example, in the Schober and Conrad (1997) study, the information provided by conversational interviewers to clarify the question was accurate (conformed to the official definitions) in 93% of the cases in which it was given. This 93% consisted of 87% in which respondents received accurate information from interviewers and provided accurate answers, and only 6% in which respondents received accurate information but provided inaccurate answers. For the 7% of cases in which interviewers provided inaccurate information, respondents were still accurate about half the time. The 7% consisted of 4% in which respondents received inaccurate information from interviewers yet provided accurate answers, and only 3% in which they received inaccurate information and answered the question inaccurately. Overall, conversational interviewers provided highly accurate information. When they did provide inaccurate information, it did not necessarily lead respondents to produce incorrect answers; in fact, respondents produced incorrect answers resulting from inaccurate information only 3% of the time.

Closer analysis of the interviewer–respondent interaction (see Conrad & Schober, 2000; Schober & Conrad, 1997) shows that it really was interviewers' deviations from standardization that led to the increases in response accuracy. It would seem that interviewer intervention improved response accuracy whether respondents had requested clarification or not (that is, even when interviewers provided the information without the respondents' having asked for it). For example, in the Schober and Conrad (1997) study, for the 64 complicated-mapping cases in which interviewers provided unsolicited help, respondents produced 55 accurate answers, an accuracy rate of 86%. In contrast, for the 11 complicated-mapping cases in which interviewers did not provide any help, respondents produced only four accurate answers, an accuracy rate of 34%. This figure is close to the 28% accuracy rate for complicated mappings in standardized interviews and suggests that when conversational interviewers do not provide clarification but behave like their standardized counterparts, response accuracy will suffer.

Across our studies, this improvement in response accuracy came at a significant cost. Conversational interviews took much longer than standardized interviews (from 80 to 300% longer, in the various studies); this was true for both straightforward and complicated mappings. Apparently, clarification just takes time. In Schober and colleagues (2004), experiment 1, the correlation between interview time and response accuracy across the five types of interviewing was .98. The more flexibility given to interviewers, the more accurately respondents answered, and the longer the interviews lasted. Practitioners will need to decide how certain they must be that respondents understand all questions as intended. If they are willing to live with some uncertainty, then it may be possible to shorten interviews while maintaining levels of response accuracy above what we observed for strictly standardized interviews.

The results of our studies should not be taken as the final word on the issue, nor should they be taken as showing definitively that conversational interviewing is always a good idea. Our studies have examined nonsensitive fact-based questions, and the results may not generalize to questions about sensitive topics or opinions in a straightforward way. Just like fact-based questions, opinion questions contain phrases with alternative possible interpretations—consider *abortion* or *approve*—and thus opinion surveys might benefit from more collaborative approaches to interviewing. But whether this can be done without influencing the opinions is unclear, especially because response accuracy for opinions cannot be validated as directly as it can for the fact-based questions in our studies. Nonetheless, O'Hara and Schober (2004) present evidence that differences in attitudes toward euthanasia are related to how respondents define the concept. One implication of this result is that presenting uniform wording does not guarantee uniform interpretation of the "attitude object," and so attitude researchers may well wish to standardize the attitude object by defining it as part of the question.

Our results also do not take into account the potential real-world costs of implementing more collaborative interviewing techniques. Beyond the potential expenses associated with increased interview length, interviewer training might have to be more intensive than it often is now. Interviewer behavior would have to be monitored even more closely to ensure that question meanings were being clarified appropriately and uniformly, without increasing interviewer variance. Far more effort would have to go into developing clear definitions for question concepts. To the extent that respondents find increased collaboration a burden, response rates could be affected.

Ultimately, the generalizability of our experimental findings on interviewing depends on the frequency of complicated mappings between questions and respondents' circumstances in real surveys; this frequency may vary from survey to survey. For the 10 questions we used in the Conrad and Schober (2000) national telephone sample, about 11% more answers changed when clarification was given (than when no clarification was given); we cannot say whether this is an accurate estimate of complicated mappings in other surveys, but it was based on actual questions from U.S. government surveys conducted with a national, representative sample. In the study by Suessbrick and colleagues (2000) mentioned earlier, respondents who answered questions about tobacco use such as "Have you smoked at least 100 cigarettes in your entire life?" exhibited a surprisingly large number of interpretations, with some respondents including only cigarettes they had finished, others including marijuana cigarettes and cigars, and others only cigarettes they had bought. This variability of interpretation was great enough that 10% of respondents changed their answers to this question when provided with a standardized definition. Because their answers determined what questions were presented in the remainder of the interview, a disturbingly large number of respondents were routed down what ultimately turned out to be the wrong questionnaire path. The Suessbrick and colleagues (2000) study was conducted in the laboratory with a convenience sample; nonetheless, the findings are consistent with other evidence (e.g., Belson, 1986) that suggests that ordinary words in survey questions are interpreted in numerous ways. Overall, our findings suggest that if complicated mappings are known to be rare, then strictly standardized techniques could lead to accurate responses at lower costs than collaborative techniques. If the complicated mappings are known to be frequent, or if (more realistically) their frequency is unknown, more collaborative techniques might be worth the increased costs their use would, no doubt, entail.

Dialogue Features in Automated Interviewing Systems

When respondents self-administer a paper-and-pencil survey or type and click answers on a Web survey, they are usually conceived of as doing something quite different than they are when they answer questions asked by a human interviewer. With self-administered questionnaires, respondents typically read rather than hear questions; they control the pacing of the interaction, and they write rather than speak their answers. Also, there is no interviewer present who might react to the respondent's answers (and potentially judge the respondent on the basis of those answers). The combination of such mode differences is no doubt what leads respondents to willingly report more socially undesirable behaviors (e.g., drug use and taboo sexual practices) in self-administered computer-based "interviews" than in actual interviews conducted by human interviewers (e.g., Tourangeau & Smith, 1996).

We propose, in contrast to the usual view, that new technologies used for presenting self-administered survey interviews fall on a continuum of anthropomorphism, with different kinds of self-administration incorporating different features of human interviewing (see Conrad & Schober, 2008). For example, textual administration steps closer to human administration when it is supplemented with audio files of computer-generated voices asking questions; computerized data collection becomes more like a human-administered interview when the text is replaced entirely with recorded human voices (see Couper, Singer, & Tourangeau, 2004). It becomes even more human-like when the interviewing system provides human-like prompts, feedback, and clarification. A self-administered questionnaire with a drawing of a human interviewer is less anthropomorphic than one with an animated, virtual interviewer whose lips move in synchrony with its speech and whose eyes blink; if the agent nods and smiles in response to answers, the system becomes yet more human-like.

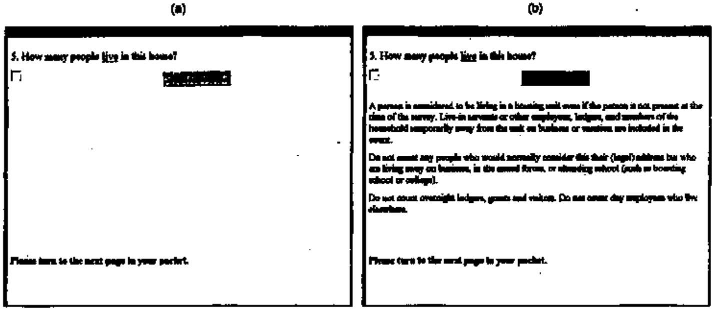
If one conceives of interviewing as involving interaction between the respondent and an interviewing agent that is either a human or a computer program, ranging from fully conversational to robotically standardized, we can begin testing which features of interview interaction lead to high-quality data and respondent satisfaction (Conrad, Schober, & Coiner, 2007). We can also begin to better understand the nature of interviewing by decomposing interviewing behaviors into their separable parts and beginning to test what happens when we add features of human interviewing to self-administered interviews.

For example, interviewers, like most conversationalists, probably make certain assumptions about the respondent's abilities as a conversationalist: They probably judge whether a particular respondent is more or less likely to need help in understanding questions and may interpret a respondent's behavior for a particular question in light of such judgments. We refer to such judgments as "respondent models." Such models may not be useful to all interviewers, but to those who are empowered to provide clarification,

that is, conversational interviewers, respondent models may help in calibrating the degree to which a respondent needs help at a particular moment in an interview.

We have implemented simple respondent models in text-based questionnaires displayed in a Web browser (Conrad, Schober, & Coiner, 2007, Experiment 2) and in a simulated speech-based system over a telephone (Ehlen, Schober, & Conrad, 2007). In the text-based system, respondents could click on a highlighted word in the question to request a definition; or, if respondents did not either click or type for more than a predetermined period of time, the system offered them a definition (see Figure 8.1). In the speech-based system, respondents could ask for clarification; or, if they exhibited spoken evidence of comprehension difficulty, such as *ums* and *uhs* or pauses (see Schober & Bloom, 2004, for evidence on which cues reliably predict misunderstanding of survey questions), the system offered them a definition. To the respondents, the speech system appeared to be automated—that is, it seemed to produce and recognize speech—but in actuality, a human experimenter played speech files in response to what respondents said in order to create the perception of automation. We developed versions of both systems that reacted differently depending on the respondent model.

When the interviewing system (whether text- or speech-based) could volunteer clarification, it interpreted respondent behavior (e.g., inactivity or silence) based either on a generic or a group-based (or "stereotype") respondent model. Under the generic model, the system treated the behavior of all respondents identically. Under the group-based model, a particular behavior (e.g., a pause of 2 seconds) was interpreted differently for respondents in different groups. The respondent attribute that we intended to model was mental quickness; a particular interval of no respondent activity or speech may signal difficulty for a quick respondent but ordinary thinking for a slower respondent. Thus help should be most useful after shorter lags for quick respondents and longer lags for slow respondents. Rather than assigning respondents to groups on the basis of their quickness, we used their age as a surrogate for quickness based on the well-known impact of cognitive aging on response time (e.g., Salthouse, 1976, 1982).



(a)

5. How many people live in this house?

□

Please turn to the next page in your packet.

(b)

5. How many people live in this house?

□

A person is considered to be living in a housing unit even if the person is not present at the time of the survey. Live-in servants or other employees, lodgers, and members of the household temporarily away from the unit on business or vacation are included in the count.

Do not count any people who would normally consider this their (legal) address but who are living away on business, in the armed forces, or attending school (such as boarding school or college).

Do not count overnight lodgers, guests and visitors. Do not count day employees who live elsewhere.

Please turn to the next page in your packet.

FIGURE 8.1. Survey question with linked definition (a) before and (b) after definition is displayed. From Conrad, Schober, and Coiner (2007). Copyright 2007 by John Wiley & Sons, Ltd. Reprinted by permission.

Therefore, under the group-based model, the system offered clarification sooner to younger respondents than to the older respondents.

With both the text-based and speech-based systems, respondents answered questions very similar to those used in the laboratory interview studies discussed earlier (Schober & Conrad, 1997; Schober et al., 2004), using either complicated or straightforward scenarios. As in the interview studies, our focus was on response accuracy for complicated scenarios. When group-based respondent models were used in either system, they improved response accuracy relative to generic response models, presumably because the clarification was better tailored to respondents' needs (see Figures 8.2 and 8.3). Irrespective of the model, enabling the system to both volunteer clarification and respond to explicit respondent requests increased response accuracy relative to cases in which respondents could obtain clarification only by explicitly requesting it. Any clarification the system provided led to greater accuracy than no clarification, which approximates what happens in standardized interviews, as well as in most current Web surveys.

Just as longer-than-normal response times may signal difficulty of some kind, quicker-than-normal response times may also reflect suboptimal performance. If a respondent answers immediately, it is unlikely that he or she has given as much thought to the task as the researchers would like. If this is the case, it may be possible to create more accurate respondent models by designing them so that the system offers help when response times are outside an optimal range, either too slow or too fast. We have dubbed this the "Goldilocks range" and used it to model respondent speech (Ehlen et al., 2007). As can be seen in Figure 8.3 (generic and stereotyped respondent models), response accuracy was better when the system could volunteer clarification based on either respondent silence or overly quick answers than when it could not provide clarification.

These findings suggest that some of what conversational interviews do can be implemented in self-administered surveys. But there are a number of caveats and questions that need to be addressed. For example, despite the apparent benefits of group-based models for response accuracy, some respondents found the system intervention unpleasant. For example, in postexperimental satisfaction questionnaires, a substantial portion of the older respondents indicated that they would prefer a human interviewer to the system that embodied a group-based respondent model, whereas they registered a preference for the computer-based system when it was built around a generic model or provided clarification only when respondents requested it by clicking. It could be that the modeling in this study is not as accurate as it might be and that better estimates of when respondents are having trouble might improve their satisfaction with the system. But the dissatisfaction of some respondents with this kind of system could also reflect a belief that respondents, not computers, should be in control of the interaction.

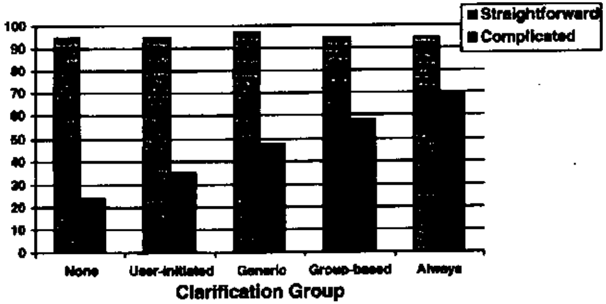


FIGURE 8.2. Response accuracy for textual question. From Conrad, Schober, and Colner (2007, Experiment 2). Copyright 2007 by John Wiley & Sons, Ltd. Reprinted by permission.

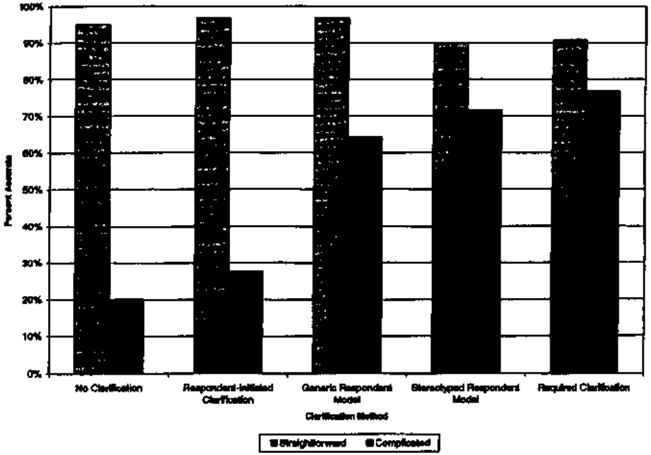


FIGURE 8.3. Response accuracy for spoken questions. From Ehlen, Schober, and Conrad (in press). Copyright by Elsevier. Reprinted by permission.

Despite the mixed impact on respondents' satisfaction, when interviewing systems offer clarification, they do so more effectively when they are based on respondent models. If interviewers and interviewing systems develop and use models of respondents, it seems likely that respondents similarly make certain assumptions about the abilities of the interviewer or interviewing system. Certainly an interface that provides clarification in conversationally savvy ways, such as adjusting its interpretation of respondent cues based on respondent characteristics, is likely to seem more human-like, or more animate. But what about embodying such animated characteristics in a virtual interviewing agent—that is, a computer-generated, graphical rendering of an interviewer—displayed in the user interface? This is sure to seem even more human. Due to the fact that animated agent technology is increasingly available, we believe it is a matter of time until survey researchers use it for collecting data. We have begun a series of experiments to explore situations in which animated agents as interviewers might help or hurt research. In order to study the impact of computerized interviewing agents that are more conversationally skilled than today's most advanced agents, we developed our agents by capturing the motions of human interviewers and using them to control the movement of graphical computer models.

This work is currently ongoing, but our experience to date suggests that the methodological concerns are quite different depending on the kinds of questions the interviewer asks. For example, when interviewing agents ask sensitive questions, do respondents assume that the agent can pass judgment on their answers as a human interviewer might? If so, they may well provide answers that are "adjusted" to be more socially desirable, as they do with human interviewers; if not, their answers will be as relatively free of social distortion as when they respond to questions that are unambiguously presented by a computer (e.g., Tourangeau & Smith, 1996). When interviewing agents ask nonsensitive, factual questions, the concerns are more focused on whether respondents believe that the interviewing agent can detect cues of response difficulty, for example, speech fluctuations and gaze aversion. If so, respondents may produce more of these cues, as we have observed in human interviews (Conrad, Schober, & Dijkstra, 2008; Schober & Bloom, 2004). Agents that are more facially realistic may encourage respondents to attribute to them the ability to detect cues of response difficulty. The agent's ability to respond to these cues, as well as to more explicit indications that help is needed, may be the most important factor, as we have observed in our studies of clarification.

An overriding issue that designers of interviewing agents will have to confront is what visual and vocal characteristics to give the agent. Human interviewers arrive on the job with certain characteristics—for example, in many organizations more interviewers are female than male. However, because the interviewer agents are animated, they can be imbued with any characteristics. It would surely be unwise to pair a female head with a male voice (see Louwerse, Graesser, Lu, & Mitchell, 2005), but other decisions are less obvious and may have a large impact on responses. For example, should a female agent wear a head scarf, or should a male agent wear a yarmulke? Should the agent match the respondent on characteristics such as age, gender, and race? If we knew the answers to such questions, we might be able to design interviewing systems that promise to make the results more credible and valid (see Conrad & Schober, 2008).

Discussion and Conclusions

We began this chapter by considering the limits of standardized interviewing as a path to comparable data across respondents, in view of the fact that different respondents can interpret the same words differently, particularly depending on how their circumstances correspond to the words. We provided evidence that allowing interviewers to choose their wording in order to make sure that respondents interpret the questions the same way and as intended can dramatically improve response accuracy, particularly when the correspondence between respondents' circumstances and question wording is ambiguous (complicated mappings). Our argument, in essence, is that standardizing meaning is more effective than standardizing wording in making data comparable. And introducing dialogue technology to self-administered surveys, as we discussed in the second part of this chapter, pushes us to rethink what counts as comparability in surveys.

To be more specific, if we think of wording versus meaning as one dimension on which comparability might be achieved, dialogue technology vastly increases the space of potential comparability. Although it is clear that the technology can help standardize respondents' interpretations by defining concepts much as human interviewers can do, it also makes it possible to tailor the interaction to respondents' abilities (e.g., mental speed), cultural conversational practices (e.g., some cultures are more tolerant of interruption than others), sensory abilities (e.g., font size or speech volume can be adjusted for the respondent), and so forth. Interviewing agents consist of many features (face, voice, clothing) that can be modified to match the respondent when desired (e.g., vocal similarity between speakers and listeners can lead to more positive ratings of speakers; Giles & Powesland, 1975) or to make them mismatch (e.g., a teenager may be less likely to inflate reports of drug use when the interviewing agent does not look like a peer). In this last case, dissimilarity could increase comparability by removing elements that lead some respondents to respond in a socially desirable way, but there may be no such effect for other respondents. Thus comparability may involve similarity to the interviewing agent for some respondents and dissimilarity for others.

The notion of comparability in standardized interviewing, we argue, deserves rethinking even in human interviews. Some characteristics of interviewers, such as vocal pitch range and race and gender, are immutable, but others are not: think of how differently warm an interviewer might be with different respondents, or how much more clarification or encouragement an interviewer might provide to different respondents. One could argue that this kind of variability is exactly what standardized interviewing intends to avoid—in the ideal, interviews would be so standardized that only one interviewer conducts them all. But is this really desirable? If we want to have truly comparable data, might it not be useful for interviewers to tailor their warmth or encouragement to the level that particular respondents need? Clearly, even having one interviewer does not necessarily standardize the stimulus to respondents. With interviewing agents, survey designers will need to make choices about what they mean by standardization and comparability and to decide which attributes to hold constant and which to allow to vary between respondents. We propose that the goal of collecting comparable data could end up requiring interviews that look rather different on the surface.

Note

1. It is widely acknowledged that interviewers' observable attributes can affect answers if interviewers differ on attributes that are relevant to the content of survey questions for which there may be more and less socially desirable answers. For example, in one study (Kane & Macaulay, 1993), female interviewers elicited more feminist responses from both men and women than did male interviewers. In another study (Hatchett & Schuman, 1975) black interviewers elicited more liberal responses to questions about racial topics than did white interviewers; for questions about nonracial topics, the answers were unrelated to the interviewer's race. We do not consider this kind of interviewer effect in the remainder of the chapter, that is, effects of the interviewer's observable characteristics; instead, we focus on the impact of interviewers' behavior on the accuracy of answers.

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 Press.

Beatty, P. (1995). Understanding the standardized/non-standardized interviewing controversy. *Journal of Official Statistics, 11*, 147–160.

Belson, W. A. (1986). *Validity in survey research*. Aldershot, UK: Cower.

Cahn, J. E., & Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. In S. E. Brennan, A. Giboin, & D. Traum (Eds.), *Psychological models of communication in collaborative systems* (pp. 25–33). Menlo Park, CA: AAAI Press.

Clark, H. H. (1991). Words, the world, and their possibilities. In G. Lockhead & J. Pomerantz (Eds.), *The Perception of Structure* (pp. 263–277). Washington, DC: American Psychological Association.

Clark, H. H. (1992). *Arenas of language use*. Chicago: University of Chicago Press.

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* (pp. 127–149). Washington, DC: American Psychological Association.

Clark, H. H., & Gerrig, R. J. (1983). Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior, 22*, 591–608.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes, 2*, 19–41.

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

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

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

Conrad, F. G., & Schober, M. F. (Eds.). (2008). *Envisioning the survey interview of the future*. New York: Wiley.

Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology, 21*(2), 165–187.

Conrad, F. G., Schober, M. F., & Dijkstra, W. (2008). Cues of communication difficulty in telephone interviews. In J. M. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. Lavrakas, M. Link, & R. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 212–230). New York: Wiley.

Couper, M. P., Singer, E., & Tourangeau, R. (2004). Does voice matter? An interactive voice response (IVR) experiment. *Journal of Official Statistics, 20*, 551–570.

Ehlen, P., Schober, M. F., & Conrad, F. G. (2007). Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Discourse Processes, 44*(3), 245–266.

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

Gibbs, R. W. (1994). *The poetics of mind*. Cambridge, UK: Cambridge University Press.

Giles, H., & Powesland, P. E. (1975). *Speech style and social evaluation*. London: Academic Press.

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

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

Louwerse, M. M., Graesser, A. C., Lu, S., & Mitchell, H. H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology, 19*, 693–704.

Maynard, D. W., & Whalen, M. R. (1995). Language, action, and social interaction. In K. Cook, G. Fine, & J. House (Eds.), *Sociological perspectives in social psychology* (pp. 149–175). Boston: Allyn & Bacon.

O'Hara, M., & Schober, M. (2004). Attitudes and comprehension of terms in opinion questions about euthanasia. *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Reddy, M. J. (1993). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed., pp. 164–201). Cambridge, UK: Cambridge University Press. (Original work published 1979)

Salthouse, T. A. (1976). Speed and age: Multiple rates of age decline. *Experimental Aging Research, 2*, 349–359.

Salthouse, T. A. (1982). *Adult cognition: An experimental psychology of human aging*. New York: Springer-Verlag.

Schober, M. F. (1998). Conversational evidence for rethinking meaning. *Social Research, 65*, 511–534.

Schober, M. F., & Bloom, J. E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse Processes, 38*, 287–308.

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., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 18*, 169–188.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85*(409), 232–253.

Suessbrick, A., Schober, M. F., & Conrad, F. G. (2000). Different respondents interpret ordinary questions quite differently. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 907–912). Alexandria, VA: American Statistical Association.

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

Traum, D. R. (1994). A computational theory of grounding in natural language conversation. Unpublished doctoral dissertation, University of Rochester, Department of Computer Science.

Handbook of Emergent Methods

EDITED BY

Sharlene Nagy Hesse-Biber

Patricia Leavy

THE GUILFORD PRESS
New York London