Journal of Official Statistics, Vol. 21, No. 2, 2005, pp. 215–231

Promoting Uniform Question Understanding in Today's and Tomorrow's Surveys

Frederick G. Conrad[1] *and Michael F. Schober*[2]

Survey respondents misunderstand questions more frequently than one might expect but, current methods for collecting data make it hard to detect and correct misunderstanding. The conventional practice has been to leave the interpretation of questions up to respondents; interviewers react to requests for clarification with nondirective probes like "Let me repeat the question." The current article reviews a research program that has explored alternatives to standardized wording, in which interviewers and web survey systems can define survey concepts as needed as a way to assure uniform comprehension across respondents. One problem is that many respondents fail to recognize that their understanding is not aligned with the survey sponsors' and so do not ask for clarification – a problem that, we argue, is more serious in the survey response task than other tasks in which information is exchanged. Using today's survey techniques (telephone and face-to-face interviews, web surveys) it is possible to increase respondents' sensitivity to their own misunderstanding, increasing their requests for clarification; and, based on respondents' verbal and visual cues of comprehension difficulty, it is possible to intervene to correct misunderstanding. This approach can be extended in surveys of the future by incorporating mature speech recognition capabilities, modeling respondent uncertainty about question meaning so that when clarification is needed it can be provided automatically, and developing interface agents when appropriate. By evaluating simulated versions of these technologies in the near term researchers will be better able exploit them as they become available.

Key words: Clarification; conversational interviewing; conceptual variability; question comprehension; standardization; interviewing techniques; mode effects; interviewer-respondent interaction.

1. Introduction

When people misunderstand questions it is not good for their answers. Yet today's techniques for collecting survey data do not make it easy to correct misunderstanding. Consider, for example, this interchange from a telephone survey, discussed in greater detail in Schober and Conrad 2002. (In this excerpt, overlapping speech is enclosed in asterisks; pauses are indicated by periods surrounded by spaces (.); lengthened sounds are indicated by colons.)

[1] University of Michigan, Institute for Social Research, P O Box 1248, Ann Arbor, MI 48106, U.S.A.
Email: fconrad@isr.umich.edu
[2] New School for Social Research, Dept. of Psychology, New York, NY 10003, U.S.A.
Email: schober@newschool.edu.
Acknowledgements: We thank the National Science Foundation (Grants SBR-97-0140, IIS-0081550, SES-0454832), the Institute for Social Research at the University of Michigan, The U.S. Bureau of Labor Statistics, The U.S. Bureau of the Census, and the Vrije Universiteit of Amsterdam.

I: And now we'd would like to ask about your employment status . did you do any work . for pay . last week
R:. Eh well . I'm still getting paid but school's out . so .
I: Okay s:o . would you say . I mean . *it's-*
R: *well*
I: it's your c*all*
R: *I g*ot paid . *for work . but I was*n't at work
I: *okay huh huh huh* hhh okay

The interviewer takes this as a "yes" answer, as evidenced by subsequent questioning. Later in the interview it turns out that, according to what the survey designers wanted to count as "work for pay," the respondent's answer should have been "no."

What went wrong here? The interviewer actually did exactly what she had been trained to do, leaving the interpretation of "work for pay" up to the respondent ("it's your call"). In current practice interviewers are typically discouraged from clarifying survey concepts, because to do so for some respondents and not others would lead to nonstandardized presentation of questions. The standardized alternative at the other extreme, providing scripted clarification for all respondents whether or not they seem to need it, is clearly undesirable; it would result in sometimes providing clarification when it is not necessary and would make the task unnecessarily burdensome. Yet not providing clarification when it is needed, as in our example, clearly can lead to measurement error.

Exactly the same dilemma arises when surveys are self-administered. Paper questionnaires offer designers no flexibility in clarifying concepts – definitions can either be presented to all respondents or none, and it seems unlikely that respondents actually read such definitions as often as they might need to. Today's web-based questionnaires in principle offer more promise for clarifying concepts on an as-needed basis, for example, by allowing respondents to click for a definition. However, this still requires respondents to recognize their need for clarification and to be willing to act on it, which, unfortunately, they often do not seem to do.

In the current article, we report results from our research program documenting the costs of leaving question interpretation up to respondents and exploring techniques to assure more uniform interpretation in both interviews and automated data collection. We will divide our discussion into two parts. First we will discuss the costs of misunderstanding questions with current methods and how misunderstanding might be reduced. Then we discuss new and potential techniques for improving people's understanding of survey questions and the quality of their answers.

The approach in our research program has been to focus on situations in which we can determine the extent to which respondents' conceptions of survey concepts match and do not match the survey designers'. We thus focus on response validity, as opposed to other valuable measures of survey quality like response rates and reliability, because we see validity – accuracy of answers – both as understudied and as the most direct determinant of survey data quality. We have set up our studies so that respondents' answers to the survey questions are informative about their conceptual alignment with the survey designers. In some studies we have probed respondents about their conceptions of the question concepts after they participate in a survey (e.g., Suessbrick, Schober, and Conrad

2000; 2005); in others we have asked respondents to answer questions on the basis of fictional scenarios that we have designed to examine response validity (e.g., Coiner, Schober, Conrad, and Ehlen 2002; Conrad and Schober 1999; Lind, Schober, and Conrad 2001; Schober and Bloom 2004; Schober and Conrad 1997; Schober, Conrad, and Fricker 2004; Schober, Conrad, Ehlen, and Fricker 2003); in still others we assess conceptual alignment by measuring response change between an interview and a reinterview (Conrad and Schober 2000) or between an interview and a self-administered questionnaire (Conrad, Schober, and Dijkstra 2004; Suessbrick, Schober, and Conrad 2000; 2005).

As we see it, interpretation of questions in survey responding is part of the larger set of issues examined throughout the social sciences about how people understand each other's terms, take each other's perspectives (or do not) and come to conceptual alignment with their conversational partners in different discourse settings. As such, our findings in the survey setting can also contribute to basic debates about the nature of conceptual alignment (see, e.g., Pickering and Garrod 2004, and the replies) and joint action more generally (e.g., Clark 1996; Schober 1998; 1999). They also contribute to the debates within the survey methods community about the merits and drawbacks of standardization (e.g., Beatty 1995; Fowler and Mangione 1990; Houtkoop-Steenstra 2000; Schaeffer 2002; Suchman and Jordan 1990; among others).

2. Clarifying Meaning in Today's Surveys

2.1. Clarification improves response accuracy

Standardized question wording is widely advocated but not strictly implemented. As Viterna and Maynard (2002) showed, the training practices and documents of twelve academically oriented survey organizations that purported to conduct standardized interviews actually vary substantially on a standardization continuum. Ten of the twelve organizations were, on average, positioned on the nonstandardized end of the continuum. For example, when respondents provide an answer that does not match any of the categories provided by the interviewer, the standardized procedure is for the interviewer to reread all of the categories (see, e.g., Fowler and Mangione 1990, pp. 39–40). However, Viterna and Maynard found that eight of the twelve organizations authorized the interviewers to determine which categories to repeat. Similarly, strictly standardized practice requires that any feedback to respondents be nonevaluative, e.g., "Thank you." Yet, a majority of organizations authorized interviewers to provide encouraging feedback such as "Well done" or "Good job."

If official policy (embodied in training documents and practices) departs from the ideals of standardized question presentation, it seems likely that in actual data collection situations interviewers will depart from strict standardization even more. Of course interviewers misspeak when attempting to read questions as worded, but they also sometimes change wording substantively (see, e.g., Fowler and Cannell 1996). Although it has typically been assumed that this will harm the accuracy of answers (because different respondents answer different questions), interviewers may actually be changing wording in order to make sure respondents understand as intended, i.e., to standardize interpretation. Dykema, Lepkowski, and Blixt (1997), for example, observed that for a

question about doctor visits, substantive changes in question wording led to more accurate answers with respect to health records. While we do not know the content of those changes, it seems they must have clarified the question authors' intentions.

In one study of survey practices at a government facility that subscribes to the philosophy and practice of standardized data collection (Schober, Conrad, and Fricker 2004, Experiment 2), interviewers varied substantially in their adherence to strict standardization, with at least one deviation occurring in an average of about 20% of the question-answer sequences. Only one of eleven interviewers followed strictest standardization to the letter. Ten of the eleven provided definitions in response to requests for clarification or asked informative follow-up questions one or more times per interview; three interviewers deviated from standardization for at least four of twelve questions in each interview, up to as many as six questions.

In contrast to what would be predicted by advocates of standardized wording, these departures from scripted question delivery led to dramatically greater response accuracy than when interviewers read only what was scripted. In this study, interviewers telephoned respondents in a laboratory, all of whom answered the same twelve questions about facts and behaviors on the basis of fictional scenarios. The scenarios were designed so that there were right and wrong answers based on official definitions of survey concepts. Depending on the particular scenario given to respondents to answer the question, half of the questions were designed to be ambiguous without clarification and half straightforward. For example, when the question asks if there have been purchases of household furniture and the scenario is a receipt for the purchase of a floor lamp, it is ambiguous whether the answer should be "yes" or "no": whether a floor lamp should count as furniture depends on how it is defined (see Figure 1 for examples of what respondents saw). It was for this kind of question – we call them complicated mappings because the correspondence between questions and what they refer to is complicated (Schober and Conrad 1997) – that providing definitions substantially improved accuracy.

The beneficial effects of clarifying question meaning are not confined only to laboratory settings where respondents answer on the basis of fictional scenarios. In a survey of 227 respondents from a U.S. national telephone sample (Conrad and Schober 2000), respondents answered ten questions about housing and purchases in a strictly standardized

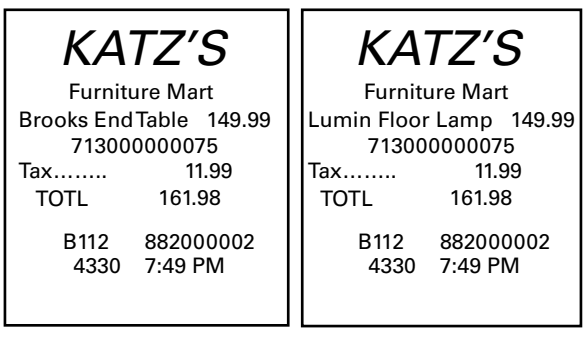

Fig. 1. (a) Straightforward scenario and (b) complicated scenario from Schober and Conrad (1997) and Schober, Conrad, and Fricker (2004)

interview where question interpretation was left entirely up to respondents. Every time they answered "yes" to a purchase question (e.g., "In the past year, have you had any purchases or expenses for moving?"), they were asked to list what those purchases had been. One week later, respondents answered the same questions again with different interviewers. This time, for half the interviews respondents could request and be given clarification if they thought they needed it, and interviewers could provide unsolicited clarification in their own words if they thought the respondents needed it.

The finding was that respondents' answers changed more often when the second interview allowed clarification (22% of answers) than when it did not (11% of answers), and that the change resulted from improved alignment between respondents' and the survey designers' understanding of the concepts in the questions. That is, the purchases that respondents who received clarification now listed were much more likely to fit the survey designers' definitions of what should count as, for example, moving.

Not only do these findings suggest that clarification improves comprehension, but they also give rough estimates, for a few survey concepts, of the extent to which respondents in the population might need clarification. Although this varied from concept to concept, respondents seemed to have life circumstances akin to the complicated mappings in our lab scenarios about 11% of the time.

2.2. When should clarification be given?

While it is clear that defining concepts for respondents can improve their understanding and thus response accuracy, the amount of improvement may depend on whether the clarification is given only after respondents request it or also when interviewers believe it is needed. The amount of improvement could also depend on whether the clarification consists of verbatim or paraphrased definitions. Schober, Conrad, and Fricker (2004, Experiment 1) compared response accuracy under strictly standardized conditions – where interviewers could not provide any clarification – and four versions of "conversational interviewing" in which interviewers were able to clarify concepts when requested by respondents (Respondent-initiated) but which differed (1) in whether the interviewers could also volunteer clarification (Mixed-initiative) and (2) in whether they could use their own words to clarify the questions (Paraphrase vs Verbatim).

Schober et al. (2004) observed that in complicated situations the more conversational flexibility afforded to interviewers the greater the improvement in response accuracy (see Figure 2). Respondents answered most accurately when interviewers could volunteer clarification (as well as providing it in response to explicit requests) and when they could do so in their own words (Mixed-initiative, Paraphrased: 87%). Note that even though interviewers were able to use their own words they provided the clarification accurately on 93% of the occasions they gave it. Accuracy was at intermediate levels when interviewers could initiate clarification or paraphrase definitions but not both (Mixed initiative, Verbatim: 66%; Respondent-initiated, Paraphrased: 55%) and just as high when the only clarification that interviewers could provide was verbatim definitions requested by respondents (Respondent-initiated, Verbatim: 59%). In contrast, when no clarification was available, accuracy was disturbingly low (Standardized interviews: 28%). By allowing

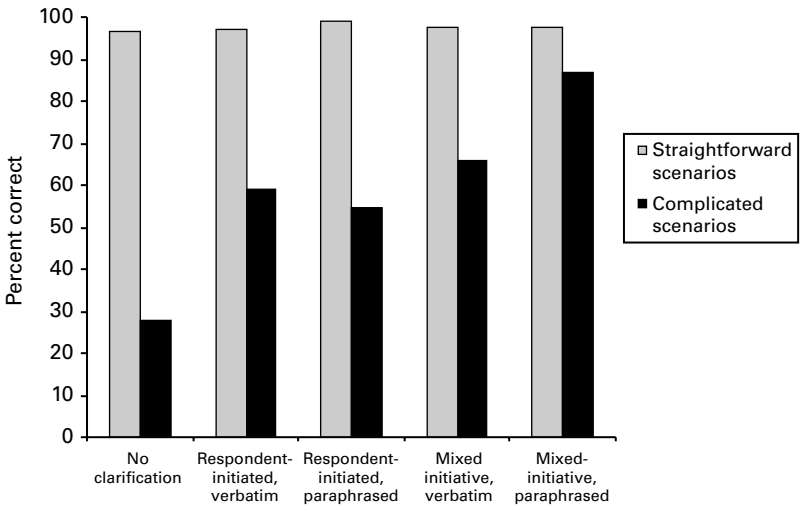

Fig. 2. Comprehension accuracy, Study 1 in Schober, Conrad, and Fricker 2004

interviewers to volunteer clarification, Mixed-initiative clarification provided important compensation for respondents' infrequent requests for definitions.

Although accuracy was high under the Mixed-initiative, Paraphrased approach, respondents requested clarification for only 47% of the complicated mappings, with interviewers initiating the balance of clarification episodes (see also Schober and Conrad 1997 for details of the verbal interaction between interviewers and respondents). Why might this be? There are at least two possibilities. First, respondents may not realize they are misinterpreting key concepts. Second, even if they are uncertain about the meaning of the question and recognize the potential benefits of obtaining clarification, they may not be willing to invest the effort needed to articulate their uncertainty to the interviewer, or they may not be willing to acknowledge uncertainty about the meaning of an ordinary term.

To gain insight into how the effort of requesting clarification affects the likelihood that respondents will do it, we carried out another laboratory study (Conrad and Schober 1999). The idea was that the effort required to request clarification should be substantially lower in a web survey, where respondents can simply click for clarification, than in a telephone interview, where respondents must think about how to formulate the spoken request for clarification. Similarly, any social obstacles to requesting clarification, such as wanting not to appear ignorant are removed in a web survey; web survey respondents have no reason to be shy about requesting clarification because there is no interviewer to potentially judge their knowledge. In the study, which used the same questions and scenarios as in Schober, Conrad, and Fricker (2004), respondents read the questions in a web browser-like application and typed their answers or selected them with the mouse. When instructed to request definitions in order be sure they understood correctly, respondents clicked for definitions in 83% of the complicated mappings – much more often than the 47% in telephone surveys using the same questions. However, when they were simply told that clarification was available and not anything about why it might be useful, they requested clarification even less often than on the telephone, clicking for only 23% of complicated mappings.

Apparently formulation difficulty is only one obstacle to requesting clarification. Respondents must also be aware that clarification could be useful.

2.3. Respondents' awareness that clarification is needed

Unfortunately, the structure of the survey task may discourage respondents from questioning their interpretation of questions – at least relative to other tasks in which information is exchanged. Consider responding in web surveys. Usually, respondents are invited by the researcher to participate, probably in the form of an email message or a letter. After they agree to participate, the web survey system asks them questions about their lives and opinions, which they answer based on their knowledge. There is little reason for the respondent to suspect that the ordinary words in the questions are intended to mean something other than what the respondent usually thinks they mean: after all, the respondent was individually invited to participate and it is reasonable to assume the words have been chosen to be interpretable by the respondent (Clark and Schober 1991). Beyond this, there is little incentive for respondents to question their interpretation because there really are no personal consequences of misunderstanding the question: if the answer is not accurate because the respondent has not grasped the intended meaning of the question, it is the researcher's problem, not the respondent's.

By way of contrast, consider a web search task. Someone (a user) in need of information on the web (say cost-of-living information to update a contract) submits a query to the system. Unlike in the web survey situation, the user initiates the search task, i.e., there is no invitation, and the user poses the query to the system (possibly by typing a string of words into Google or clicking on a link), not the other way around as in the survey task where the system queries the user. The information in the search task resides in the system, not in the user's head, again in contrast to the web survey task. The system returns the information requested in the query and the user than carries out some action(s) with the retrieved information in order to fulfill the overall goal of the task. Correctly interpreting the words on the web page is the user's concern; misunderstanding will interfere with achieving the goal. These differences in task structure lead to the prediction that web survey respondents should request clarification less often than comparable web users engaged in a search task, because survey respondents have less reason to question their interpretation.

This prediction was tested by Schober, Conrad, Ehlen, and Fricker (2003). Laboratory participants were asked to carry out either a web survey or web search task. In both tasks, respondents used information from fictional scenarios similar to those used in the studies described earlier, for example, a floor plan of a housing unit depicting two rooms labeled "bedroom" and one room labeled "den." In the complicated version of the scenario, the den functioned as a bedroom, despite its original design as a den, so that the number of bedrooms to be reported depended on whether or not the den should be counted as a bedroom. Participants could obtain a definition by clicking on the word "bedroom" in either task. The relevant survey question asked simply how many bedrooms the house contained. The comparable web search task involved retrieving a table of rental prices in order to determine the monthly rent of the unit. Because the table listed rent according to the number of bedrooms, participants had to be clear on the number of bedrooms because misinterpreting the bedroom would lead to the

wrong conclusion about the rental price. Based on the logic just presented, we expected more requests for clarification in the search task than in the survey task.

Indeed those engaged in the web search task requested clarification for complicated mappings twice as often as those in the web survey task, on 47.7% versus 23.4% of occasions, supporting the idea that by virtue of its structure, the survey response task does not promote respondents' awareness that they may not understand as intended. Participants' open-ended comments in a debriefing questionnaire were consistent with this explanation of the differences in rates of requesting clarification. Thirty-nine percent of web search participants reported that the questions were "tricky," required attention to detail, or were definitionally uncertain, e.g., "At the beginning of the study I presumed to know the definitions of some of the terms. It wasn't until part way through the study did I come to the understanding that my def. may be different." In contrast only 12.5% of web search participants voiced similar sentiments. More typical was the following comment, revealing little suspicion that word meaning might not have been what it seemed: "This was interesting that I did not have to think hard to complete the task. I enjoyed responding to the questions."

Can survey respondents be made more aware that clarification might be helpful? One promising approach to increasing respondents' awareness is to alter the question wording so as to imply that the key concept in the survey question might have alternate interpretations. One can do this by including a piece of the definition for a key concept along with the question – but not the entire definition. For example, one could alter the question "How many people live in your home?" by adding "Live-in servants and other employees are included in the count." Even if a respondent's circumstances do not include live-in servants, this might suggest that the notion of who lives in one's home might have exceptions and complications.

Lind, Schober, and Conrad (2001) tested this idea in a laboratory study. Respondents read questions in a web browser and registered their answers either by typing or selecting options with the mouse. They could request clarification by clicking the mouse on highlighted text, in response to which the system would display the definition below the question. Respondents either received original question wordings or altered wordings, and the altered wordings either contained components of definitions relevant to their fictional circumstances or components of definitions that were irrelevant. For example, a respondent whose scenario described a family with a child away at college would either be asked the original question *How many people live in this house?*, a version altered to include a relevant component of the definition (*Do not count 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)*), or a version with an added irrelevant component of the definition (*Live-in servants and other employees are included in the count*).

The results showed that when a directly relevant piece of the definition was included, respondents did not click for the full definition any more (21.4% of the time) than when they read the question as originally worded (25.0% of the time). But when an irrelevant component of the definition for residence was included with the question, respondents clicked for clarification reliably more often, 42.7% of the time. It was as if the inclusion of irrelevant definitional content seemed to suggest to respondents that the concepts were more complicated than they might otherwise have realized, piquing their curiosity enough

to request the full definition. And, of course, when respondents received the appropriate clarification, their response accuracy improved.

2.4. Degrees of conceptual misalignment between respondents and survey designers

The distinction between relevant and irrelevant features of a concept, and the idea that definitions have multiple features, suggests that respondents and survey designers can align their understanding to varying degrees. This implies that we need a more nuanced view of how respondents and survey designers might be aligned and misaligned, because the features they are misaligned on might or might not be relevant to the circumstances about which respondents are answering. So if the respondent is asked about any moving expenses, it does not matter how differently the respondent conceives of moving expenses than the survey designers if the respondent has not engaged in any activity that could conceivably count as moving; the respondent's answer will, correctly, be "no" despite the possibility of multiple misalignments. On the other hand, a respondent whose conception is aligned with the survey designers' on every dimension but the one relevant to his or her circumstances will still provide an inaccurate answer. For example, the respondent might correctly realize that *moving expenses* includes charges for packing, freight and storage and does not include charges for U.S. postal delivery service but incorrectly believe (be misaligned on the feature) that *expenses for moving* includes charges for do-it-yourself moving, like trailer rental. So misalignment should sometimes lead to misunderstanding and response error, but other times to adequate understanding and accurate responding.

Suessbrick, Schober, and Conrad (2005) demonstrate this in a study of concepts about smoking. 125 respondents answered the questions in the Tobacco Supplement to the U.S. Current Population Survey, a survey which has been administered every two years since 1948. Questions concerned behaviors like "Have you ever tried cigars, pipes, chewing tobacco or snuff?" and opinions like "In restaurants do you think that smoking should be allowed in all places at all times, allowed in some places at some times, or not allowed at all?" Respondents answered about their own lives but in a lab setting; after the telephone interview they completed a post-survey conceptualization questionnaire and then a self-administered reinterview. The conceptualization questionnaire, a multiple choice questionnaire concerning possible meanings of the survey concepts, allowed us to assess degrees of conceptual alignment with survey designers in a sample of New Yorkers. The self-administered reinterview, in which they answered the same survey questions as in the interview but accompanied by standard definitions, allowed us to assess how often different kinds of misalignment led to inaccurate answers, as measured by corrections to those answers in the reinterview (which we refer to as "unreliable responding").

Among the various findings, the most relevant here is that over 59% of the time respondents were conceptually misaligned with the survey designers on at least one feature of the definitions. But this conceptual misalignment did not necessarily lead to inaccurate responding; respondents answered unreliably only 45% of the time they were misaligned – often enough to worry about, but certainly not constantly. Nonetheless, the misalignment can be quite serious, as in the degree to which responses to the very first question, "Have you smoked at least 100 cigarettes in your entire life?", turned out to be based on varying interpretations of what counts as smoking cigarettes (with some people

including marijuana, pipes, cigars and others excluding them; with some people excluding cigarettes they had only taken a puff from or that they had not bought themselves, etc.). Answering this question with nonuniform interpretation led 10% of respondents down the wrong path (smoker or nonsmoker) in the survey, and would substantially affect estimates ot the prevalence of smoking in the U.S.

2.5. *Cues that respondents need clarification*

A repeating theme in the studies described thus far is that respondents do not ask for clarification as much as they need to even when the interviewer or survey system encourages it. For example, as mentioned earlier, Conrad and Schober 1999 observed far fewer requests for clarification from respondents who were simply told that definitions were available than that they were essential to understanding the question as it was intended.

Closer analyses of the human-computer and human-human interaction data across studies suggest that even when respondents do not explicitly request clarification, they can provide other behavioral and linguistic cues that they are having trouble answering a question. They can take a very long time to answer in a web or telephone survey. They can provide disfluent answers, *um*ming and *uh*ing, restarting, and repeating themselves. They can report their circumstances ("I bought a floor lamp") rather than answering with the required "yes" or "no," leaving the interpretation of their answer up to the interviewer.

As evidence that these cues can reliably indicate respondents' need for clarification, Schober and Bloom (2004) examined the strictly standardized and mixed-initiative paraphrased clarification telephone interviews from Schober and Conrad (1997). As these were interviews based on fictional scenarios, it was possible to know when respondents were answering about straightforward circumstances and when they were answering about complicated circumstances. If we look at all the question-answer sequences where respondents did not request clarification explicitly, the first utterance respondents made after each survey question was asked did indeed contain reliable cues for whether clarification was needed. In particular, respondents were more likely to *um* or *uh*, to restart their utterances, to report their circumstances, and to pause longer when answering about complicated mappings. Interestingly, they tended to provide more such cues in interviews that allowed clarification, as if they were at some level aware that interviewers might be able to use the cues to judge when to offer unsolicited clarification. It is hard to know whether these cues are intended to be communicative; in any case, they do seem to provide evidence about need for clarification.

Just as telephone interviews afford more cues of respondents' need for clarification than text-only web surveys, face-to-face interviews can potentially provide more, because respondents and interviewers can see each other. The Schober and Bloom (2004) study focused on telephone interviews where, by definition, interviewers are not privy to any visual cues of uncertainty that respondents might display – at least with today's ordinary telephone technology. This raises the question of how sensitive respondents are to the relative richness of cues afforded by the mode of data collection. More specifically, do respondents compensate for the absence of visual cues in telephone interviews by displaying more cues of uncertainty in their speech than they do in face-to-face interviews? If respondents are sensitive to which cues interviewers can and cannot perceive, are they

also sensitive to whether interviewers can and cannot act on those cues? In other words, do they display more such cues in conversational than standardized interviews because they recognize (at some level) that conversational interviewers are licensed to respond to uncertainty cues with clarification in a way that standardized interviewers are not?

Conrad, Schober, and Dijkstra (2004) tested this in a laboratory study in which Dutch respondents were asked about their own lives in either conversational or standardized interviews that were conducted over the telephone or face-to-face. After the interview, respondents self-administered a paper questionnaire that included the interview questions accompanied by definitions of the relevant concepts. Thus, if respondents changed their answers between the interview and the post-interview questionnaire, the change could be attributed to a change in their understanding brought about by reading the definition in the questionnaire.

Conrad et al. (2004) focused their analysis on a question about membership in the Dutch institution *vereiningen* or registered clubs: "I would now like to ask you some questions about your membership in clubs. Can you list all the clubs in which you are personally a member?" This type of question, which requires respondents to list their answers, is a particularly good candidate for conversational interviews because interviewers can help respondents evaluate each club they list for compliance with the definition (see also Conrad and Schober 2000). Indeed, answers changed more in standardized than conversational interviews for this question, suggesting that clarification during the (conversational) interview was beneficial to respondents' understanding and the accuracy of their answers. However, there were no differences due to mode (telephone versus face-to-face). Why might this be, particularly given the extra richness in potential cues of uncertainty afforded by face-to-face interviews?

Part of the answer lies in respondents' greater disfluency over the telephone. In particular, they produced more *um*s and *uh*s per word on the telephone than face-to-face, as if they recognized that the interviewers could not see them on the telephone. What then are the visual cues available only in face-to-face for which telephone respondents may have been compensating? One such potential cue is respondents' gaze aversion, that is, their tendency to look away from the interviewer while answering. Increased gaze aversion has been associated with increased difficulty in answering questions (Glenberg, Schroeder, and Robinson 1998) and is attributed to the respondents' attempt to avoid the distraction that is almost certainly brought about by looking at the questioner's face. The critical issue in the Conrad et al. (2004) study was whether respondents looked away from the interviewer while answering the question more in conversational than standardized interviews.

In fact respondents did look away for larger percentages of time when answering questions posed by conversational than standardized interviewers: in cases where their answers later proved reliable, respondents looked away 15.4 percent of the time while answering in conversational interviews, as compared with 4.3 percent of the time in standardized interviews. More tellingly, in cases where their answers later proved unreliable they looked away 28.3 percent of the time in conversational interviews (versus 0 percent of the time for standardized interviews, where there was no chance they could get clarification). These data suggest that respondents were sensitive to whether the interviewers could provide clarification in response to a visual behavior. However, conversational interviewers did not provide more clarification in response to this behavior,

despite glancing at respondents at least once during 80% of the turns in which they looked away. One explanation is that conversational interviewers simply had not been instructed to treat such cues as indications of respondent uncertainty and that with appropriate training they could provide more and better-timed clarification. Another possibility is that interviewers were so focused on looking at their laptop screens that they were not sufficiently aware of respondents' gaze aversion to use it as a cue of need for clarification.

3. Clarifying Meaning in Tomorrow's Surveys

As we learn more about the kinds of cues that respondents present that indicate their need for clarification, we can begin not only to speculate but also to explore experimentally how surveys of the future might best be implemented.

3.1. Interviews that use today's technologies

For surveys of the future that use today's technologies and media (face-to-face, telephone, web), we propose that interviews that promote clarification are vital to accurate survey measurement. Of course, much more needs to be known about when and how clarification makes a difference. The studies by Conrad and Schober (2000) and Suessbrick, Schober, and Conrad (2000) demonstrate that complicated mappings between respondents' circumstances and survey concepts lead to misconceptions often enough to compromise overall data quality. Yet Suessbrick, Schober, and Conrad (2005) demonstrate that not all misconceptions necessarily produce incorrect answers, i.e., if the respondent and survey sponsors' concepts are misaligned on features not directly relevant to the answer, response accuracy may be unaffected by misunderstanding. So clarifying all aspects of survey concepts may not be necessary. This is underscored when one considers that clarifying concepts takes time. Conrad and Schober (2000) observed that conversational interviews were 80% longer than standardized interviews and Schober and Conrad (1997) observed a three-fold increase in interview duration between conversational and standardized interviews (in a laboratory situation where 50% of questions involved complicated mappings). So the benefits of improved alignment must be weighed against the costs of bringing this about.

We believe the costs are worthwhile. Alarming as the increase in interview duration might be, the benefits cannot be denied. Schober, Conrad, and Fricker (2004) found that time and accuracy were highly correlated ($r = .97$) across their five interviewing conditions; each additional minute of interviewing (clarifying concepts) produced a 7% increase in accuracy for the twelve survey questions they asked. The promise of conversational techniques in interviews and web-based data collection, it seems to us, lies in better diagnosis of when respondents will benefit from clarification and when they will use it, i.e., recognize its value. If clarification is provided just under these circumstances, the cost-benefit ratio can be made much more favorable.

We advocate continued use of laboratory and field methods to determine when it is worth going to the additional expense and effort that appropriate clarification techniques will require. As we see it, the studies described here have only begun to tackle a set of serious questions about clarification in interviewing. For example, our studies have focused on surveys about facts and behaviors rather than on surveys about opinions and

attitudes. The evidence thus far (O'Hara and Schober 2004; Suessbrick, Schober, and Conrad 2000, 2005) is that interpretive variability is at least as great for terms in attitude questions as it is for terms in questions about facts and behaviors. This leads to the worrisome prospect that if the interpretation of terms in attitude questions is left up to respondents, some percentage of what is being measured may be respondents' interpretations of the terms in the questions rather than their attitudes. Yet while it is surely desirable that respondents interpret attitude objects uniformly, it is unclear whether clarification could be provided that does not bias reported attitudes.

Similarly, with the exception of Suessbrick, Schober, and Conrad (2000; 2005), our studies have tended to involve sets of questions excerpted from larger surveys, and they have used relatively small populations of test respondents. While respondents in a national telephone sample seemed no less likely to participate in interviews with clarification than without (Conrad and Schober 2000), it is not yet known whether response rates would remain stable if overly time-consuming clarification were deployed in a longer interview.

Another arena requiring investigation is the extent to which interpretive variability differs for different survey questions and different populations. Although we have found substantial variability in every domain we have examined, the set of studies reported here has been conducted in only a fraction of possible domains. We suspect that interpretive variability is more the norm than the exception. But we also have evidence that some words in survey questions are interpreted less uniformly than others, and we do not yet know the guiding principles for predicting when a question is likely to be interpreted problematically – when the frequency of complicated mappings in the general population is high.

We believe that these gaps in our knowledge are worth filling, because ignoring the problems of question misunderstanding will not make them go away.

3.2. Interviews that use up-and-coming technologies

As communications technologies rapidly develop and shift the ways the population communicates and is contactable, the media in which surveys are likely to take place will evolve. Although survey researchers have been conservative in adopting new technologies in recent years, they will have no choice but to adapt in order to continue measuring the public's behavior and opinions. We believe clarification techniques can be fruitfully extended in several directions to keep pace with evolving technologies and media. We now turn to two of these.

3.3. Speech interfaces

Web collection of survey data is almost exclusively done via desktop interfaces, i.e., questions are displayed textually and respondents answer by typing or selecting options with the mouse. Because speech contains cues of potential respondent uncertainty (Conrad, Schober, and Dijkstra 2004; Schober and Bloom 2004), automated clarification could potentially be delivered more effectively if the data collection system had access to these cues. Speech recognition is used to a limited degree in current survey telephone interfaces (so-called Integrated Voice Recognition (IVR)), but this is used in limited vocabulary applications, e.g., "yes"/"no" or numerical responses in which paralinguistic information is not taken into account. But more sophisticated speech recognition is coming

of age and can be deployed (whether on the web or in more conventional telephone interfaces) to detect comprehension problems of the sorts we have been discussing.

Our group (Bloom 1999; Ehlen 2005; Ehlen, Schober, and Conrad 2005; Schober, Conrad, and Bloom 2000) has explored these issues with simulated speech interfaces in which respondents believe they are interacting with a computer but are in fact listening to speech files presented by a human experimenter and responding to the experimenter over the telephone. Respondents displayed a range of verbal cues (restarts, confirmations, *um*s and *uh*s, unfilled pauses, etc.) more often when their answers were based on complicated rather than straightforward situations. In the Ehlen (2005) study, latency before speaking was the cue most predictive of the need for clarification: if it was very brief or long, respondents benefited more from system-initiated clarification than if the latency was of intermediate duration. While the automated delivery of clarification was simulated by a human using an automated tool, a fully automated system could monitor respondents' speech for such cues without attentional lapses to which a human interviewer would undoubtedly fall prey.

3.4. User (respondent) modeling

In the Ehlen (2005) study, the system provided clarification based on the respondent's speech behavior. Time before responding was interpreted as an indication of conceptual alignment – if the respondent answered too quickly or too slowly it was taken as evidence that the respondent was either overconfident (when too quick) or confused or struggling (when too slow), presumably much as sensitive and savvy human interviewers model respondents. This is a simple model of respondents' comprehension state.

Ehlen refined this by treating the latency differently depending on respondents' age (similar to the study by Coiner, Schober, Conrad, and Ehlen 2002, in which the interpretation of inactivity with a desktop interface was calibrated by age group). Respondents answered more accurately when the system presented clarification based on its model of them, whether the model was generic or based on age group – in fact, as accurately as with constant clarification. This suggests that generic and group-based respondent modeling may be a way of providing clarification when needed while avoiding clarification when it is not needed. Clearly, the next step in this line of research is to tailor the models to individual respondents, based perhaps on their performance on several test questions at the outset of the data collection session.

3.5. Envisioning the future of interviewing

It is not a great leap to imagine that interviewing interfaces of the future will become more and more sophisticated, embodying more and more of the features of human interviewers, and generally blurring the line between interviewing and self-administration. Existing technologies already show the way: in AutoTutor (e.g., Graesser, Moreno, Marineau, Adcock, Olney, and Person 2003; Graesser, Wiemer-Hastings, Wiemer-Hastings, and Kreuz 1999), a talking head presents questions with appropriate facial expressions, prompts for more information when necessary, and manages mixed-initiative dialogue smoothly. The data show that talking heads in the interface improve student learning by increasing students' willingness to ask for help, as well as increasing students' satisfaction with the interaction and their likelihood of continuing the session.

The kinds of data we have been collecting on what leads to valid answers will be useful in thinking through what features future interviewing systems should include, and when they might be most usefully deployed to encourage participation and survey completion as well as the most accurate answers. Obviously a talking-head system could be deployed in the most strictly standardized ways, leaving the interpretation of questions up to respondents, or it could be deployed with increasing levels of dialogue sophistication so as to present generic or more individually tailored clarification. A system could be designed to recognize users' temporal, paralinguistic and facial cues as they answer questions. Of course, much is unknown about what sorts of features would lead to improved data quality and response rates. One could imagine that for sensitive or embarrassing questions, for example, respondents might answer more accurately with a less anthropomorphic agent that feels more anonymous. For some question domains or respondent populations, one might want to turn on or off various features.

As these technologies become more commonplace and the population becomes more used to interacting with them, there will be pressure for survey designers to develop them for surveys. Our proposal is that systematic study of response accuracy in simulations of future systems can help survey designers decide what kinds of interviewing systems are worth developing. The body of evidence described here, and the sort of laboratory and field experimentation it represents, may help us navigate the uncharted territories of survey interviewing in new communications media.

4. References

Beatty, P. (1995). Understanding the Standardized/Nonstandardized Interviewing Controversy. Journal of Official Statistics, 11, 147–160.

Bloom, J.E. (1999). Linguistic Markers of Respondent Uncertainty During Computer-administered Survey Interviews. Unpublished dissertation, New School for Social Research.

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

Clark, H.H. and Schober, M.F. (1991). Asking Questions and Influencing Answers. In J.M.Tanur (ed.), Questions About Questions: Inquiries into the Cognitive Bases of Surveys, 15–48. New York: Russell Sage Foundation.

Coiner, T.F., Schober, M.F., Conrad, F.G., and Ehlen, P. (2002). Assessing Respondents' Need for Clarification in Web Surveys Using Age-based User Modeling. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Conrad, F.G. and Schober, M.F. (1999). A Conversational Approach to Text-based Computer-administered Questionnaires. Proceedings of the 3rd International Conference on Survey and Statistical Computing, 91–101. Chesham, UK: Association for Survey Computing.

Conrad, F.G. and Schober, M.F. (2000). Clarifying Question Meaning in a Household Telephone Survey. Public Opinion Quarterly, 64, 1–28.

Conrad, F.G., Schober, M.F., and Dijkstra, W. (2004). Nonverbal Cues of Respondents' Need for Clarification in Survey Interviews. Proceedings of the American Statistical

Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Dykema, J., Lepkowski, J.M., and Blixt, S. (1997). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), Survey Measurement and Process Quality, 287–310. New York: John Wiley.

Ehlen, P. (2005). Proceedings of the Symposium on Dialogue Modelling and Generation. On occasion of the 15th Annual meeting of the Society for Text and Discourse, Vrije Universiteit, Amsterdam.

Ehlen, P., Schober, M.F., and Conrad, F.G., (2005). Should Computer-based Interviews Model Survey Respondents' Speech to Decide When to Clarify Terms? Paper presented at the Annual Meeting of the Society for Text and Discourse, Amsterdam Netherlands.

Fowler, Jr, F.J. and Cannell, C.F. (1996). Using Behavioral Coding to Identify Cognitive Problems with Survey Questions. In N. Schwarz and S. Sudman (eds), Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, 15–36. San Francisco, CA: Jossey-Bass.

Fowler, F.J. and Mangione, T.W. (1990). Standardized Survey Interviewing: Minimizing Interviewer-related Error. Newbury Park, CA: Sage Publications, Inc.

Glenberg, A.M., Schroeder, J.L., and Robinson, D.A. (1998). Averting the Gaze Disengages the Environment and Facilitates Remembering. Memory and Cognition, 26, 651–658.

Graesser, A.C., Moreno, K., Marineau, J., Adcock, A., Olney, A., and Person, N. (2003). AutoTutor Improves Deep Learning of Computer Literacy: Is It the Dialog or the Talking Head? In U. Hoppe, F. Verdejo, and J. Kay (eds), Proceedings of Artificial Intelligence in Education, 47–54. Amsterdam: IOS Press.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R. and the TRG (1999). Auto Tutor: A Simulation of a Human Tutor. Journal of Cognitive Systems Research, 1, 35–51.

Houtkoop-Steenstra, H. (2000). Interaction and the Standardized Survey Interview: The Living Questionnaire. Cambridge, UK: Cambridge University Press.

Lind, L.H., Schober, M.F., and Conrad, F.G. (2001). Clarifying Question Meaning in a Web-based Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

O'Hara, M. and Schober, M.F. (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.

Pickering, M.J. and Garrod, S. (2004). Toward a Mechanistic Psychology of Dialogue. Behavioral and Brain Sciences, 27, 169–226.

Schaeffer, N.C. (2002). Conversation with a Purpose–or Conversation? Interaction in the Standardized Interview. In D. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds), Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, 95–123. New York: John Wiley.

Schober, M.F. (1998). Conversational Evidence for Rethinking Meaning. Social Research, Special Issue: Conversation, 65, 511–534.

Schober, M.F. (1999). Making Sense of Questions: An Interactional Approach. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (eds), Cognition and Survey Research, 77–93. New York: John Wiley and Sons.

Schober, M.F. and Bloom, J.E. (2004). Discourse Cues that Respondents Have Misunderstood Survey Questions. Discourse Processes, 38, 287–308.

Schober, M.F. and Conrad, F.G. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? Public Opinion Quarterly, 61, 576–602.

Schober, M.F. and Conrad, F.G. (2002). A Collaborative View of Standardized Survey Interviews. In D. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds), Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, 67–94. New York: John Wiley and Sons.

Schober, M.F., Conrad, F.G., and Bloom, J.E. (2000). Clarifying Word Meanings in Computer-administered Survey Interviews. In L.R. Gleitman and A.K. Joshi (eds), Proceedings of the 22nd Annual Conference of the Cognitive Science Society, 447–452. Mahwah, NJ: Lawrence Erlbaum Associates.

Schober, M.F., Conrad, F.G., Ehlen, P., and Fricker, S.S. (2003). How Web Surveys Differ from Other Kinds of User Interfaces. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Schober, M.F., Conrad, F.G., Ehlen, P., Lind, L.H., and Coiner, T.F. (2003). Initiative and Clarification in Web-based Surveys. Papers from the 2003 AAAI Spring Symposium "Natural Language Generation in Spoken and Written Dialogue," Technical Report SS-03–06, 125–132. Menlo Park, CA: AAAI Press.

Schober, M.F., Conrad, F.G., and Fricker, S.S. (2004). Misunderstanding Standardized Language in Research Interviews. Applied Cognitive Psychology, 18, 169–188.

Suchman, L. and Jordan, B. (1990). Interactional Troubles in Face-to-face Survey Interviews. Journal of the American Statistical Association, 85, 232–253.

Suessbrick, A.L., Schober, M.F., and Conrad, F.G. (2000). Different Respondents Interpret Ordinary Questions Quite Differently. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Suessbrick, A.L., Schober, M.F., and Conrad, F.G. (2005). When Do Respondent Misconceptions Lead to Survey Response Error? Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Viterna, J.S. and Maynard, D.W. (2002). How Uniform Is Standardization? Variation Within and Across Survey Research Centers Regarding Protocols for Interviewing. In D. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds), Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, 365–397. New York: John Wiley.

Received April 2005
Revised June 2005